Fast and powerful statistical method for context-specific QTL mapping in multi-context genomic studies

Andrew Lu*[1], Mike Thompson[2], M Grace Gordon[6], Andy Dahl[7], Chun Jimmie Ye[8], Noah Zaitlen[3,4,5], and Brunilda Balliu*[3]

[1]UCLA-Caltech Medical Scientist Training Program, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
[2]Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA
[3]Department of Computational Medicine, University of California Los Angeles, Los Angeles, CA, USA
[4]Department of Human Genetics, University of California Los Angeles, Los Angeles, CA, USA
[5]Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA
[6]Biological and Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, CA, USA
[7]Section of Genetic Medicine, University of Chicago, Chicago, IL, USA
[8]Division of Rheumatology, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA
*Corresponding author; email: andrewluchun@ucla.edu; bballiu@ucla.edu

Abstract

Recent studies suggest that context-specific eQTLs underlie genetic risk factors for complex diseases. However, methods for identifying them are still nascent, limiting their comprehensive characterization and downstream interpretation of disease-associated variants. Here, we introduce FastGxC, a method to efficiently and powerfully map context-specific eQTLs by leveraging the correlation structure of multi-context studies. We first show via simulations that FastGxC is orders of magnitude more powerful and computationally efficient than previous approaches, making previously year-long computations possible in minutes. We next apply FastGxC to bulk multi-tissue and single-cell RNA-seq data sets to produce the most comprehensive tissue- and cell-type-specific eQTL maps to date. We then validate these maps by establishing that context-specific eQTLs are enriched in corresponding functional genomic annotations. Finally, we examine the relationship between context-specific eQTLs and human disease and show that FastGxC context-specific eQTLs provide a three-fold increase in precision to identify relevant tissues and cell types for GWAS variants than standard eQTLs. In summary, FastGxC enables the construction of context-specific eQTL maps that can be used to understand the context-specific gene regulatory mechanisms underlying complex human diseases.

1 Introduction

Genetic variants associated with complex disease reside mainly in the non-coding component of the genome, leading to the natural hypothesis that they act through transcriptional regulation [1]. Large-scale multi-context expression quantitative trait loci (eQTL) studies have demonstrated extensive sharing of eQTL effects across contexts, such as tissues and cell types [2–5], environmental stimulation [6], advanced aging [7], etc. For example, characterization of cis eQTLs across 49 human tissues in the Genotype-Tissue Expression (GTEx) project has revealed cis eQTLs for 95% of protein-coding genes in at least one tissue [2, 3] and sharing of 85% of eQTLs across tissues [5]. This pervasive sharing complicates the mechanistic understanding of complex trait associations and prioritization of the disease-relevant contexts for eQTLs.

Interestingly, eQTLs often exhibit complex patterns of context-specific effects, wherein a variant can regulate, to a different degree, the expression of a gene across many contexts [5]. Characterization of these variants will allow a better understanding of gene regulation and disease etiology. Indeed, mounting evidence suggests that genetic variants underlying disease associations are often context-specific [8–16]. For example, the Immune Variation project identified eQTLs in monocyte-derived dendritic cells and human CD4+ T lymphocytes with different effects in response to *in vitro* stimulation and polarization [13, 17]. These previously unknown, immune state-specific eQTLs strongly overlapped autoimmune disease-associated variants [6, 18, 19]. Similarly, [20] mapped eQTLs during differentiation of induced pluripotent stem cells to cardiomyocytes to identify eQTLs that change over time. These dynamic eQTLs were enriched for genes with roles in myogenesis and dilated cardiomyopathy.

To identify context-specific eQTLs (sp-eQTLs) while constraining experimental heterogeneity and reducing costs, studies often gather multiple samples across contexts for the same donors [3, 17, 21, 22]. Linear mixed models (LMMs) are a natural analysis choice for such studies [23–25] because they model the intra-individual correlation inherent across repeated samples and directly identify sp-eQTLs by testing for the significance of the genotype-by-context (GxC) interaction term. However, these LMMs are computationally infeasible for eQTL studies. Hence, researchers instead rely on simple linear models with a GxC (LM-GxC) term [9, 20] or context-by-context (CxC) eQTL mapping, followed by post hoc examination of summary statistics to distinguish shared and context-specific eQTL effects [2, 3]. While relatively fast, these approaches are significantly underpowered because they do not leverage intra-individual correlation in multi-context studies like GTEx (Figure S2) and single-cell RNA-Seq data [26]. Additionally, many rely on downstream, ad hoc definitions of context-specific and shared genetic effects that are based on subjective, manually selected thresholds of effect size differences between contexts [5] or presence-absence of effects in different contexts [3, 8, 27, 28]. These definitions can have a large impact on context-specific eQTL mapping by under- or over-counting sharing of effects across contexts. These shortcomings have limited characterization of sp-eQTLs and downstream interpretation of disease-associated variants.

To address these limitations, we introduce FastGxC, a novel method that leverages the correlation structure of multi-context studies to efficiently and powerfully map sp-eQTLs. In brief, Fast-

GxC decomposes the phenotype of interest per individual into context-shared and context-specific components and estimates genetic effects on these factors separately using simple linear models. We prove through analytical derivation and empirical examination that FastGxC shared and context-specific effect size estimates are a reparametrization of the CxC and LMM-GxC estimates. FastGxC has several key advantages over previous methods. First, by removing the intra-individual correlation, it naturally adjusts for background noise unrelated to the context of interest, e.g., sex, age, and sequencing batch [7, 29, 30]. Second, it uses ultra-fast implementations of linear regression models specifically designed for eQTL mapping [31]. Third, it directly maps sp-eQTLs without the need for post hoc analyses or arbitrary thresholds. Fourth, it provides both global and marginal tests for sp-eQTLs. The global test identifies variants with eQTL effect size heterogeneity across contexts while the marginal tests identify the context(s) driving this heterogeneity. FastGxC output integrates naturally with recent methods developed to improve the statistical power of CxC eQTL mapping, such as mash [5], sn_spMF [4], and Meta-Tissue [32, 33]. FastGxC is broadly generalizable to any continuous phenotype, e.g., bulk or single-cell gene expression [3, 34], protein and metabolic measurements [21, 22], and DNA methylation levels [35], measured across different contexts, e.g., tissues and cell types [36–38], environmental perturbations [17, 19], developmental stage [35], aging, [7, 22], and differentiation state [20].

We first show in simulations that FastGxC is as powerful as the LMM-GxC but orders of magnitude faster. Both approaches are orders of magnitude more powerful than a heterogeneity test based on CxC estimates and LM-GxC in the presence of intra-individual correlation. We next applied FastGxC to multi-tissue RNA-Seq data from the GTEx Consortium [3] and peripheral blood single-cell RNA-Seq data from CLUES, an in-house 234 person cohort (see accompanying manuscript), to produce the most comprehensive tissue- and cell-type-specific eQTL map to date across 49 tissues and eight peripheral blood cell types. We validate these maps by establishing enrichment of sp-eQTLs in corresponding functional genomic annotations. Finally, we examine the relationship between FastGxC sp-eQTLs and human disease and show that they provide a three-fold increase in precision to identify relevant contexts for GWAS variants across 138 complex traits compared to standard eQTLs. In summary, FastGxC enables the construction of context-specific eQTL maps that can be used to understand the context-specific gene regulatory mechanisms

underlying complex human diseases.

2 Results

FastGxC method overview. We illustrate FastGxC using tissues as the contexts (Figure 1A) but the method can be applied to different contexts, e.g., cell types and environmental stimuli. Briefly, for each individual, FastGxC decomposes the gene expression across C contexts into one context-shared component and C context-specific components (Figure 1A - Decomposition step). Next, FastGxC identifies contexts-shared and contexts-specific eQTLs (sh-eQTL and sp-eQTL) by estimating genetic effects on the context-shared expression component and each of the contexts-specific components (Figure 1A - eQTL mapping step). FastGxC then performs a global test for context-specific eQTLs which identifies variants with significant eQTL effect size heterogeneity across contexts. Last, to identify the context(s) driving this heterogeneity, FastGxC performs C marginal tests for the significance of each of the context-specific eQTLs.

More formally, let E_{ic} be the observed expression of a gene for individual i ($i = 1, \ldots, I$) in context c ($c = 1, \ldots, C$). FastGxC first decomposes E_{ic} into an offset term, a context-shared component, and a context-specific component [39], i.e.

$$E_{ic} = E_{..} + \underbrace{(E_{i.} - E_{..})}_{E_i^{sh}} + \underbrace{(E_{ic} - E_{i.})}_{E_{ic}^{ts}} \quad (1)$$

where $E_{..} = \left(\sum_{i=1}^{I} \sum_{c=1}^{C} E_{ic}\right) / (I \times C)$ is the average expression of the gene, computed over all I individuals and all C contexts, and $E_{i.} = \left(\sum_{c=1}^{C} E_{ic}\right) / C$ is the average expression of the gene for individual i, computed over all contexts. In (1), $E_{..}$ is a term that is constant across individuals and contexts for each gene, E_i^{sh} is the context-shared expression component for individual i and is constant across contexts for each gene and individual, and E_{ic}^{ts} is the context-c-specific expression component for individual i.

Next, FastGxC estimates one shared and C context-specific cis genetic effects by regressing the genotypes on each component using ultra fast implementations of fixed-effect linear regression models [31], i.e.,

$$
\begin{aligned}
E_i^{sh} &= \alpha^{sh} + \beta^{sh} G_i + \varepsilon_i^{sh}, \\
E_{i1}^{ts} &= \alpha_1^{ts} + \beta_1^{ts} G_i + \varepsilon_{i1}^{ts}, \\
&\vdots \\
E_{iC}^{ts} &= \alpha_C^{ts} + \beta_C^{ts} G_i + \varepsilon_{iC}^{ts},
\end{aligned}
$$

where $\alpha^{sh}, \alpha_1^{ts}, \ldots, \alpha_C^{ts}$ are offsets. G_i is the genotype of individual i, coded as number of minor alleles, and β^{sh}, $\beta_1^{ts}, \ldots, \beta_C^{ts}$ are the genetic effects on the shared and each of the context-specific expression components. Finally, $\varepsilon_{i1}^{ts}, \varepsilon_{i1}^{ts}, \ldots, \varepsilon_{iC}^{ts}$ are each normally distributed residual errors with mean zero and variances $\sigma_{sh}^2, \sigma_{ts,1}^2, \ldots, \sigma_{ts,C}^2$.

FastGxC defines a *shared-eQTL* (sh-eQTL) as a variant with a statistically significant effect on the shared component, i.e. β^{sh}, and a *context-specific eQTL* (sp-eQTL) as a variant with at least one statistically significant genetic effect on the context-specific expression components, i.e. $\beta_1^{ts}, \ldots, \beta_C^{ts}$. The later test is performed using Simes's procedure [40]. In addition, FastGxC defines a *sp-eQTL in context c* as a variant with a statistically significant genetic effect on the context-c-specific expression component. Figure 1B illustrates different patterns of sh-eQTL and sp-eQTL effects. Notably, FastGxC shared and context-specific eQTL effect size estimates are a reparametrization of the CxC and L(M)M-GxC estimates (S3E). Full details of the analytical derivation and relationship to previous approaches are provided in the Methods and Supplementary Text.

FastGxC is more powerful and orders of magnitude faster than existing methods in simulation studies. We evaluate the global and marginal type I error rates and power of FastGxC in a series of simulations and compare its performance to a CxC-based test of eQTL effect size heterogeneity and the LM-GxC and LMM-GxC approaches. In order to obtain global estimates of type I error rate and power for each method, we test the global null hypothesis of no heterogeneity of genetic effects across contexts. Specifically, for the CxC-based approach, we fit a linear model for each context c ($E_{ic} = \alpha_c + \beta_c G_i + \varepsilon_{ic}$), and test the null hypothesis of no eQTL effect heterogeneity across contexts ($H_0 : \beta_1 = \ldots = \beta_C = 0$) using the heterogeneity statistic Q from a random-

effects meta-analysis as implemented in the `meta` R package [41]. For the LM-GxC approach, we fit one linear model with a genotype-by-context interaction term ($E_{ic} = \alpha + \beta_1 G_i + \sum_{c=2}^{C} \gamma_c K_{ic} + \sum_{c=2}^{C} \delta_c G_i \times K_{ic} + \varepsilon_{ic}$) and test the null hypothesis of no genotype-by-context interaction effects ($H_0 : \delta_2 = \ldots = \delta_C = 0$) using the likelihood ratio test. For the LMM-GxC approach, we fit one linear random effects model with a genotype-by-context interaction term ($E_{ic} = u_i + \alpha + \beta_1 G_i + \sum_{c=2}^{C} \gamma_c K_{ic} + \sum_{c=2}^{C} \delta_c G_i \times K_{ic} + \varepsilon_{ic}$ $u_i \sim N(0, \sigma_i^2)$) and test the same null hypothesis as the LM-GxE model. Finally, for FastGxC, we test the presence of at least one context-specific effect using Simes's method for combining p-values [40]. To assess the ability of FastGxC to identify the heterogeneous context(s), we also obtain marginal estimates of type I error rate and power within each context.

We simulate 10,000 data sets for each scenario. We assume that, in each scenario, gene expression is measured in five contexts for 100 individuals. In each scenario, we vary the amount of intra-individual correlation, i.e. correlation of gene expression across contexts within individuals, from zero, i.e. no intra-individual correlation, to 0.8, i.e. high intra-individual correlation. We set the mean of the gene expression in each context to one. Genotypes for each individual were simulated using a binomial distribution with a minor allele frequency of 0.2. Under the null hypothesis of no genetic effect heterogeneity, the effect of the genotype is the same in each context (similar to toy example illustrated in Figure 1B - second panel), i.e., $\beta_j = 0.1$ for $j = 1 : 5$. We simulated two scenarios under the alternative hypothesis of genetic effect heterogeneity. In the first scenario ("single-context heterogeneity"), one context had different genetic effects from the other four contexts (Figure 1B - third panel), i.e., $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.1$ and $\beta_5 = 0.4$, and in the second scenario ("extensive heterogeneity"), every context had a different genetic effect from all other contexts (Figure 1B - fourth panel), i.e., $\beta_j = 0.j$ for $j = 1 : 5$.

Under the null hypothesis of no genetic effect heterogeneity, FastGxC and LMM-GxC maintain a 5% type I error rate both at the global (Figure 2A) and marginal (Figure S3A) level, regardless of the amount of intra-individual correlation. As expected, the CxC-based and LM-GxC approaches, which do not model the intra-individual correlation, become more conservative with increasing intra-individual correlation (Figure 2A). Under the alternative hypotheses of genetic effect heterogeneity, FastGxC is as powerful as LMM-GxC in both the single-context (Figure 2B) and extensive heterogeneity (Figure S3B) scenarios, and both methods become more powerful as the level of intra-

individual correlation increases. As expected from their performance under the null scenario, the CxC and LM-GxC approaches lose power in the presence of intra-individual correlation (Figures 2B and S3B). In addition, FastGxC correctly identified the context(s) that drive the heterogeneity in both the single-context (Figure 2C) and extensive heterogeneity (Figure S3C) scenarios.

To benchmark the computational costs of running FastGxC compared to the other approaches, we simulated phenotype and genotype data as above. To obtain practical run-times, we used study parameters from GTEx, the largest multi-context eQTL study to-date with approximately 50 contexts and an average of 250 individuals per context, while varying the number of tests performed (Figure 2D). When extrapolated to mapping cis-eQTLs in the entire GTEx dataset, i.e. approximately 200M tests for 25K genes and 3M SNPs, we found that LMM-GxC and LM-GxC would finish in approximately 30 years and 10 months, respectively, while CxC and FastGxC achieved equivalent results in under one minute (average run time in 100 iterations). At 1,000 individuals, FastGxC continues to be efficient (five minutes for all tests) while LMM-GxC would take upwards of 500 years (Figure S3D).

FastGxC produces a high-resolution map of tissue-specific and tissue-shared eQTLs in GTEx. We applied FastGxC to GTEx v8 RNA-seq data [3] to decompose the expression in each tissue into a tissue-shared and 49 tissue-specific components. To assess the ability of FastGxC to remove gene expression background noise, we correlated technical and biological covariates with the first ten principal components (PCs) from the original gene expression data and the decomposed tissue-shared and tissue-specific expression data (Figure 3A). As expected, the largest sources of variation in the original gene expression data, as captured by the top 10 PCs, were highly correlated to biological and technical variables such as donor sex, age, ethnicity, and cohort [7, 30, 42]. The impact of many of these sources of variation is absent in the FastGxC tissue-specific expression components, i.e., the top ten PCs from the tissue-specific expression are not correlated to variables that do not change within an individual, e.g. sex, age, and genotype PCs. These results suggest that FastGxC effectively reduces background noise inherent in gene expression data by removing the intra-individual correlation between tissues transcriptomes from the same individual.

We next mapped cis eQTLs on each of these components, providing a high resolution map

of tissue-shared and tissue-specific eQTLs (sh-eQTLs and sp-eQTLs) (Table S1). We discovered a total of 20,947 sh-eGenes, i.e. genes with at least one sh-eQTL (60.7% of tested genes; hierarchical FDR (hFDR) $\leq 5\%$) and an average of 1,620 sp-eGenes, i.e., genes with at least one sp-eQTL, per tissue (6% tested genes; hFDR $\leq 5\%$). In addition, we discovered 7,671,697 sh-eQTLs and between 9,998 (kidney cortex) and 1,008,063 (testis) sp-eQTLs within each tissue, totaling 11,656,197 sp-eQTLs across 49 tissues (Figure 3B and S1; hFDR $\leq 5\%$). Compared to the standard CxC analysis, FastGxC discovered an additional 700 eGenes, consistent with the power increase observed in the simulations (Figure S4A). Of these additional FastGxC discoveries, 60% are sh-eGenes and the remaining 40% are sp-eGenes.

We then sought to understand the sharing and specificity of FastGxC-mapped eQTLs. We found that the majority of sp-eQTLs are discovered in only a few tissues indicating that, for the majority of sp-eQTLs, few tissues drive the heterogeneity (Figure 3C). In addition, sp-eQTLs found in more biologically-distinct tissues such as testis (16%), make up the largest proportion of the sp-eQTLs that are unique to a single tissue (Figure 3C). Across tissues, most variants (85.9-97.5%) with tissue-specific eQTL effects have also shared eQTL effects (Figure 3D and S5), suggesting that most tissue-specific effects manifest within the shared effect loci and would be missed by approaches that define context-specificity by presence or absence of significant eQTL effects in each context rather than differences in sizes of eQTL effects.

Additionally, we show that sp-eQTL effect sizes are correlated between groups of biologically related tissues, e.g., sp-eQTL are shared among 13 brain, two heart (left ventricular and atrial appendage), two artery (tibial and aorta), two esophagus (muscularis and gastro-esophageal junction), three adipose (visceral, subcutaneous, and breast), and two intestine tissues (Figure 4 - right triangle). This result is consistent with the previously reported high correlation of eQTL effects between groups of biologically related tissues from the CxC approach [5]. Yet, while FastGxC sp-eQTL effect sizes show little to no correlation outside groups of biologically related tissues, CxC effect sizes show widespread correlation across all tissues regardless of biological relationships (Figure 4 - left triangle). This again demonstrates that FastGxC is able to disentangle tissue specific effects from shared effects.

Tissue-specific eQTLs are enriched in functional genomic features from their matched tissues. To validate FastGxC sh-eQTL and sp-eQTLs and understand the functional differences between variants with sh-eQTL and sp-eQTL effects, we performed enrichment analysis for genomic elements using variants with sp-eQTL but no sh-eQTL effects ("sp-eQTL only") and variants with sh-eQTL but no sp-eQTL effects ("sh-eQTL only"), compared to a random subset of minor allele frequency (MAF)-matched non-eQTL variants (Figure 5A). Sp-eQTL only variants are enriched (OR=1.06, p-value = 1.16×10^{-5}) while sh-eQTL only variants are depleted (OR=0.98, p-value = 2.87×10^{-2}) within enhancers (FDR $\leq$ 5%; Figure 5A). In addition, sh-eQTL only variants show stronger enrichment within promoters, compared to sp-eQTL only variants ($OR_{sh} = 1.14$ versus $OR_{ts} = 1.04$; p-value = 3×10^{-7}). These results are consistent with previous observations that variants with tissue-specific effects are more enriched in genomic elements that confer tissue specificity to gene expression, such as enhancers, while variants with tissue-shared effects are more common within promoters [43].

In order to understand how eQTL variants mapped by the CxC approach are functionally different than FastGxC eQTL variants, we performed enrichment analysis for genomic elements using variants that are only discovered by CxC ("CxC only") or FastGxC ("FastGxC only") (Figure 5A). Compared to CxC only variants, the FastGxC-only variants are significantly enriched (FDR $\leq$ 5%) in more genomic features and often more strongly enriched in key genomic elements such as promoter-flanking regions ($OR_{FastGxC} = 1.16$ versus $OR_{CxC} = 1.08$; p-value for OR difference= 6.4×10^{-9}) and introns ($OR_{FastGxC} = 1.05$ versus $OR_{CxC} = 1$; p-value= 2.3×10^{-10}). Additionally, FastGxC only eQTLs are significantly enriched in enhancers (OR=1.05, p-value = 2.1×10^{-3}), while CxC only eQTLs are not (OR=1.02, p-value = 1.8×10^{-1}). These results suggested that eQTLs only discovered by FastGxC and not CxC are more likely to reside in functional regions.

As chromatin and TF-binding architectures are strongly tissue-specific [44], they serve as important avenues to validate FastGxC mapped sp-eQTLs and quantify the functional differences between eQTLs mapped by FastGxC and CxC. We performed enrichment analysis of variants with FastGxC sp-eQTL and CxC eQTL effects in a single tissue in open chromatin of several ENCODE tissues. Of the 54 pairs of correctly-matched tissues, FastGxC single-tissue sp-eQTL variants are enriched in their matched ENCODE tissue more often than CxC single-tissue eQTL variants, i.e.

54% (29/54) versus 30% (16/54) of the time (McNemar test, p-value = 1.95×10^{-3}; Figure 5B). FastGxC variants are also, on average, more strongly enriched in their matched open-chromatin regions, compared to CxC variants ($OR_{FastGxC} = 1.37$ versus $OR_{CxC} = 1.18$ average across matched tissues; Paired t-test, p-value = 9.17×10^{-5}). Furthermore, we observed widespread enrichment in open chromatin for FastGxC and CxC variants with eQTL effects specific to tissues with cell-types ubiquitously found across human tissues, e.g. skeletal muscle, breast, and whole blood [45, 46].

We next performed enrichment analysis of the same sets of variants as above in the predicted, tissue-specific TF binding sites (TFBS) [47] (Figure 5C). In line with results from the chromatin accessibility data, FastGxC single-tissue sp-eQTL variants are more often enriched in their matched tissue-specific TFBS than CxC single-tissue variants, i.e. 53% (16/30) versus 17% (5/30) of the time, respectively (McNemar test, p-value = 2.6×10^{-3}; Figure 5C). In addition, FastGxC single-tissue sp-eQTL variants are, on average, more strongly enriched compared to CxC ($OR_{FastGxC} = 1.53$ versus $OR_{CxC} = 1.28$ average across matched tissues; Paired t-test, p-value = 1.5×10^{-3}).

Together these results demonstrate that the tissue-specific components better capture the underlying molecular contexts - both tissue-specific chromatin accessibility and TF binding sites - of their matched tissues than the CxC approach.

FastGxC uncovers novel and biologically relevant eQTLs that enhance our understanding of how genetic effects are shared and divergent across tissues. To provide insight into patterns of sharing and specificity of eQTL effects revealed by FastGxC, we discuss a few individual examples (Figure 6).

First, we examine *CBS*, a gene which encodes the enzyme cystathionine beta-synthase that catalyzes the rate-limiting step of the transsulfuration pathway [48, 49] (Figure 6A). This pathway acts ubiquitously across many cell-types to perform diverse and important biological functions such as protein synthesis and methylation [50]. Indeed, eQTL effect size estimates from CxC are significant in 48 individual tissues (hFDR $\leq$ 5%), suggesting a universal, shared mechanism of genetic regulation. FastGxC crystallizes this shared mechanism by identifying a single sh-eQTL and no sp-eQTLs (hFDR $\leq$ 5%).

Second, we show an eQTL for *SIGLEC14*, an immune cell surface receptor of the immunoglob-

ulin superfamily involved in the innate immune response [51] (Figure 6B). Similar to the *CBS* example, there seems to be a sharing of genetic effects across GTEx tissues which could lead one to conclude that this genetic effect is invariant across the body. Yet, when we explicitly model this sharing with FastGxC, a sp-eQTL effect in whole blood emerges, indicating that, while *SIGLEC14* is under a universal tissue-shared genetic regulation, there is importantly also a blood-specific regulatory mechanism that is consistent with the known role of *SIGLEC14* in immunity.

Finally, we discuss the genetic regulation of *LDHC*, which encodes the testis-specific enzyme lactate dehydrogenase C, the first testis-specific enzyme discovered in male germ cells [52] (Figure 6C). We found that *LDHC* exhibits a strong positive eQTL effect in all tissues except the testis for which the eQTL effect is in the opposite direction. This lone effect becomes very apparent when sp-eQTLs are examined with FastGxC. To the best of our knowledge, this is the first time that testis-specific genetic regulation, in addition to testis-specific expression, is reported for this gene, suggesting that tissue-specificity can be regulated at multiple biological levels.

We present additional examples that illustrate the power of FastGxC to map context-specific eQTL effects in Figure S6.

Tissue-specific eQTLs identify putatively causal tissues of complex traits. One of the primary goals for mapping QTLs is to find the molecular link between genetic variants and their associated diseases. As such, we next explored whether FastGxC results can lead to better understanding of the regulatory mechanisms and contexts in which these mechanisms operate in complex human diseases. Specifically, we extracted an independent set of trait-associated variants from 138 mapped traits in the NHGRI-EBI GWAS catalog [53]. We followed the protocol of the GTEx consortium and used expert curation to identify the most likely relevant tissue(s) (Table S2) [3]. We tested FastGxC sh-eQTL and sp-eQTL variants for enrichment in these sets, compared to a random and equal sized set of MAF-matched non-eQTL variants. We compare these enrichment results to ones based on variants with standard CxC eQTL effects in each tissue (Table S2).

FastGxC sh-eQTL and sp-eQTLs provide a three-fold increase in precision to identify the disease-relevant tissue(s) and a two-fold improvement in their rank compared to standard CxC eQTLs (Figure 7A). In addition, CxC eQTLs prioritize 22 of the 49 tissues tested per trait (me-

dian across traits), likely due to the large amount of tissue-sharing of CxC eQTL effects (Figure 4). By contrast, FastGxC prioritizes only five tissues per trait with a similar recall rate as CxC. This difference suggests that modeling the extensive sharing of eQTL effects across tissues has the potential to improve our ability to localize GWAS associations to a smaller subset of putatively causal tissues.

Across the board, the FastGxC enrichment patterns recapitulate known trait-tissues associations (Figure 7B, hFDR$\leq$ 5%). For example, in breast carcinoma, the tissue with the highest enrichment according to FastGxC is breast mammary tissue (OR = 5.0, P-value = 3.2×10^{-4}). On the other hand, for standard eQTLs mapped by CxC, the strongest enrichment was for EBV-transformed lymphocytes while breast mammary tissue (OR = 2.24, p-value = 7.5×10^{-4}) was the 25th most enriched tissue. In lung adenocarcinoma, the most common type of lung cancer, CxC finds significant associations in 22 tissues, many seemingly unrelated to lung physiology (lung OR = 2.83, 18th strongest enrichment of 22 tissues, p-value = 1.6×10^{-3}), while FastGxC only finds significant associations in lung (OR = 5.67, p-value = 2.6×10^{-3}) and nerve tibial (OR = 20, p-value = 2.1×10^{-5}). Interestingly, in the non tissue-specific cancer trait, we found that for FastGxC shared eQTLs showed the strongest enrichment, consistent with the idea that this trait reflects shared process across all tissues. This improved tissue resolution was also seen in non-cancer traits. For example, in coronary artery disease, CxC finds significant associations in 43 of the 49 tested tissues, while FastGxC finds only 16 and the top tissues are almost all cardiovascular-relevant, i.e. coronary and aortic artery, heart left ventricle and atrial appendage, skin, muscle, and average tissue.

Taken together, we demonstrate that FastGxC leads to improved resolution for localizing known tissue-trait associations. This result suggests that utilizing FastGxC to map context-specific eQTLs has the potential to discover novel links between contexts and diseases, and critically generate testable hypothesis for downstream experimental validation.

Cell-type-specific eQTLs are enriched for variants associated with immune-related complex traits. Single-cell RNA-seq eQTL studies provide an ideal setting for the application of FastGxC because the same donor contributes cells across almost all known cell types, leading to considerable intra-individual correlation (Figure 5A of accompanying manuscript). In addition, for

cases in which eQTLs from a complex tissue, e.g. whole blood, are enriched for disease-associated variants, single cell data provide an opportunity to examine the underlying cell types from this complex mixture. To that end, we applied FastGxC to the CLUES study, a cohort with single-cell RNA-Seq data in eight peripheral blood mononuclear cell (PBMC) types from 234 individuals (see accompanying manuscript). We identified 1,025 and 1,223 genes with at least one shared and at least one cell-type-specific cis eQTL, respectively (hFDR$\leq$ 5%). We extensively characterized these cell-type-specific eQTLs and showed that FastGxC cell-type-specific eQTLs for each cell type were significantly and specifically enriched for regions of chromatin accessibility in the same or closely related cell types (see accompanying manuscript).

We next tested for enrichment of FastGxC shared and cell-type-specific eQTLs in sets of trait-associated variants from 59 immune-related traits in the GWAS catalog (Figure 7C). We compare these results to enrichment results from CxC eQTLs mapped in the same single-cell data set, as well as enrichment results from GTEx bulk CxC whole-blood and FastGxC whole-blood sp-eQTLs. Variants with cell-type-specific eQTL effects in the single-cell PBMC (scPBMC) data are enriched for disease-associated variants of nine immune-related traits ($hFDR \leq 5\%$). For example, variants with eQTL effects specific to conventional and plasmacytoid dendritic cells are enriched for allergic rhinitis-associated variants, consistent with the crucial role of dendritic cells in the development and maintenance of rhinitis [54]. In addition, variants with eQTL effects specific to B and CD4+ T cells are enriched for rheumatoid arthritis-associated variants [26]. We observed a large overlap in the traits that were enriched for FastGxC and CxC mapped eQTLs, including the two examples highlighted above.

The rapid adaptation of single-cell technologies in the past few years has provided an unprecedented opportunity to dissect genetic regulatory mechanisms in granular cell types. In particular, we found that the allergy trait is enriched for single-cell eQTLs in plasmacytoid dendritic cells, celiac disease is enriched for natural killer cell-specific eQTLs, and chronic lymphocytic leukemia is enriched for eQTLs effects in and specific to several cell types ($hFDR \leq 5\%$). Critically, none of these trait enrichments were detected in the GTEx bulk data($hFDR \leq 5\%$). We foresee that the increase in single-cell experiment sample sizes, which will necessarily come with decreasing materials and sequencing costs, will expand the ability of FastGxC to map single-cell context-specific eQTLs.

3 Discussion

We developed FastGxC, a novel statistical method to efficiently and powerfully map context-specific eQTLs by leveraging the correlation structure of multi-context studies. We showed via simulations that FastGxC is as powerful as the state-of-the-art LMM-GxC method while orders of magnitude faster. We applied FastGxC to bulk multi-tissue and single-cell RNA-seq data sets and identified over 11 million tissue-specific and 280 thousand cell-type-specific eQTLs. Most context-specific effects manifest within loci with context-shared effects, highlighting the importance of defining context-specificity by effect size heterogeneity rather than the presence or absence of significant eQTL effects in each context. In addition, we found that tissue-specific eQTLs are shared mostly between groups of biologically related tissues and are more enriched in genomic elements that confer tissue specificity to gene expression, e.g., tissue-specific regions of open chromatin, providing further evidence of their validity. Finally, we found that context-specific eQTLs provide increased precision for identifying disease-relevant tissues across 138 complex traits, confirming their utility in understanding the context-specific gene regulatory mechanisms underlying complex human diseases.

While FastGxC is the first efficient method to leverage intra-individual correlation for identifying context-specific regulatory effects, several statistical methods using other techniques have been developed in recent years [4, 5, 32, 33, 55]. Most of these methods use matrix factorization of eQTL statistics to build data-driven priors that capture the underlying tissue-shared and tissue-specific architecture in eQTLs across tissues [4, 5, 55]. These flexible priors provide a considerable increase in power to map (context-specific) eQTLs compare to CxC eQTL mapping. However, they require extensive tuning of model hyper-parameters, making them computationally challenging for multi-context studies and complicating the interpretation of sharing and specificity of eQTLs across contexts. Interestingly, these methods are complementary to FastGxC as FastGxC output integrates naturally with these methods. The joint approaches may further increase the statistical power to map context-specific eQTLs as well as bypass the need for post hoc use of arbitrary cutoffs. Another recent work [9], use tests for interactions with inferred cell type proportion to identify interaction QTLs (iQTLs). This approach may also benefit from modelling intra-individual correlation, but can not be integrated with FastGxC directly as it requires a different mixed model.

FastGxC has several limitations. First, as done in previous work [3], we select the most relevant tissues for disease using experts in the field. However, the complete set of causal tissues is unknown, and rankings may change as we discover novel biology for each trait. Second, while the global test for context-specific eQTLs is always well-defined, the marginal tests for identify the contexts driving this heterogeneity are not, e.g., when every pair of contexts shows eQTL heterogeneity. However, we find that the marginal tests work well in practice, especially when only a few contexts drive this heterogeneity [4]. Third, FastGxC is limited to continuous phenotypes and discrete contexts. While there are natural LMM to apply outside of these situations, they are computationally inefficient. However, recent work in approximate algorithms may produce a solution [56]. Fourth, the current FastGxC method uses a decomposition with a single component shared across all contexts. It is straightforward to extend FastGxC when additional sharing exists across a subset of contexts, e.g., brain tissues in GTEx, by performing a hierarchical decomposition. Fifth, we define context-specificity as deviations of eQTL effects in each context from the effect in the average context. When, instead, deviations from the eQTL effect in a baseline context are of interest, e.g., when studying eQTL effects across time or differentiation states, it is straightforward to modify the decomposition step of FastGxC appropriately. Finally, relating context-specific eQTLs to GWAS variants is imperfect due to LD. Multi-context genomic colocalization approaches may improve the resolution of causal variants [57].

In conclusion, we show that accounting for the intra-individual correlation and extensive sharing of eQTLs across contexts reveals context-specific eQTLs that can aid downstream interpretation of disease-associated variants. Moreover, we demonstrate the advantage of defining context specificity by the heterogeneity of effect sizes rather than heuristic definitions based on subjective P-value thresholds. In the coming years, we believe that the application of FastGxC in the increasing number of multi-context bulk and single-cell RNA-Seq studies holds enormous potential to broaden our understanding of the context-specific gene regulatory mechanisms underlying complex human diseases.

Acknowledgements N.Z. is supported by NIH grants R01HG006399, R01CA227237, R01CA227-466, R01ES029929, R01MH122688, U01HG009080, R01HG011345, R35GM133531, R01HL155024,

R01MH125252, DoD grant W81XWH-16-2-0018, and the Chan Zuckerberg Science Initiative. C.J.Y. is supported by the NIH grants R01AR071522, R01AI136972, R01HG011239, and the Chan Zuckerberg Science Initiative, and is an investigator at the Chan Zuckerberg Biohub and a member of the Parker Institute for Cancer Immunotherapy (PICI). The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Author contributions B.B. conceived of the project and developed the statistical methods. A.L. and B.B. implemented the comparisons with simulated data. A.L., B.B., M.T., and M.G.G. performed the analyses of the GTEx and CLUES data and additional analyses. B.B. and A.L. implemented the software. A.L. and B.B. wrote the manuscript, with significant input from N.Z., C.J.Y., A.D., M.G.G., and M.T. A.L. and B.B. prepared the online code and data resources.

Software Availability We provide free access to the software at `https://github.com/BrunildaBalliu/FastGxC`. Due to size limitations, the map of shared and context-specific eQTLs for all GTEx tissues and all CLUES PBMCs is available upon request.

Competing interests C.J.Y. is a Scientific Advisory Board member for and hold equity in Related Sciences and ImmunAI, a consultant for and hold equity in Maze Therapeutics, and a consultant for TReX Bio. C.J.Y. has received research support from Chan Zuckerberg Initiative, Chan Zuckerberg Biohub, and Genentech.

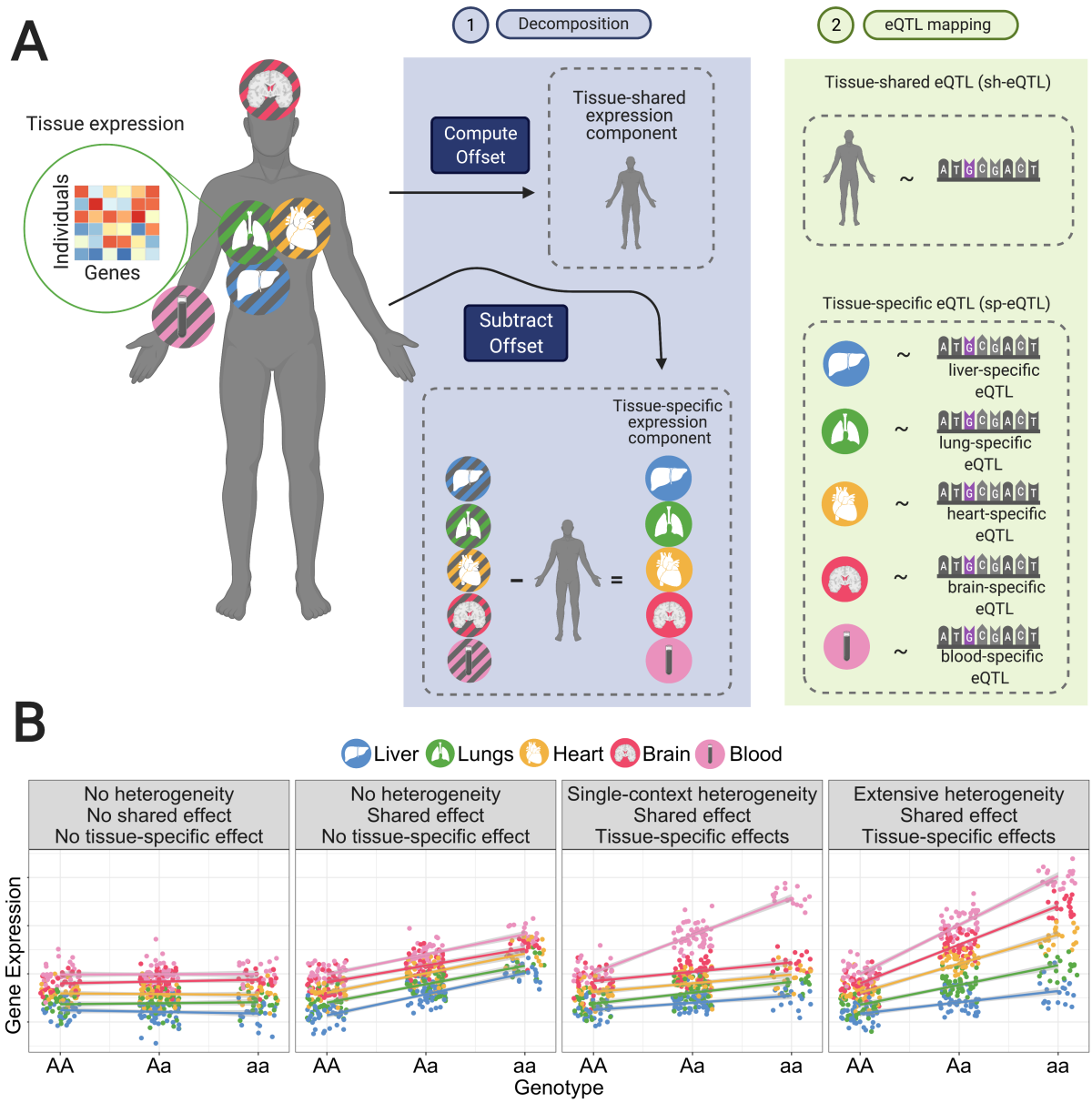


Figure 1. The FastGxC method and examples of sp-eQTLs. A. Overview of FastGxC method. FastGxC decomposes the gene expression of an individual into a context-shared and context-specific components (step 1) and estimates both the shared eQTL (sh-eQTL) effect across contexts and context-specific eQTL effects in each context by regressing the genotypes on each of these components (step 2). **B.** Toy examples of examples of sp-eQTLs. Y axis represents simulated gene expression levels, x axis lists the genotypes of a candidate eQTL, color indicates tissue. The first example corresponds to a scenario with no eQTLs in any tissue and, thus, no sh-eQTL or sp-eQTLs. The second example illustrates a scenario with equal eQTL effects in all tissues, corresponding to a scenario with a sh-eQTL but no sp-eQTLs in any tissue. The third and forth example corresponds to scenarios with both sh-eQTL and sp-eQTL effects in which a single context (blood) or multiple contexts drive the effect size heterogeneity.

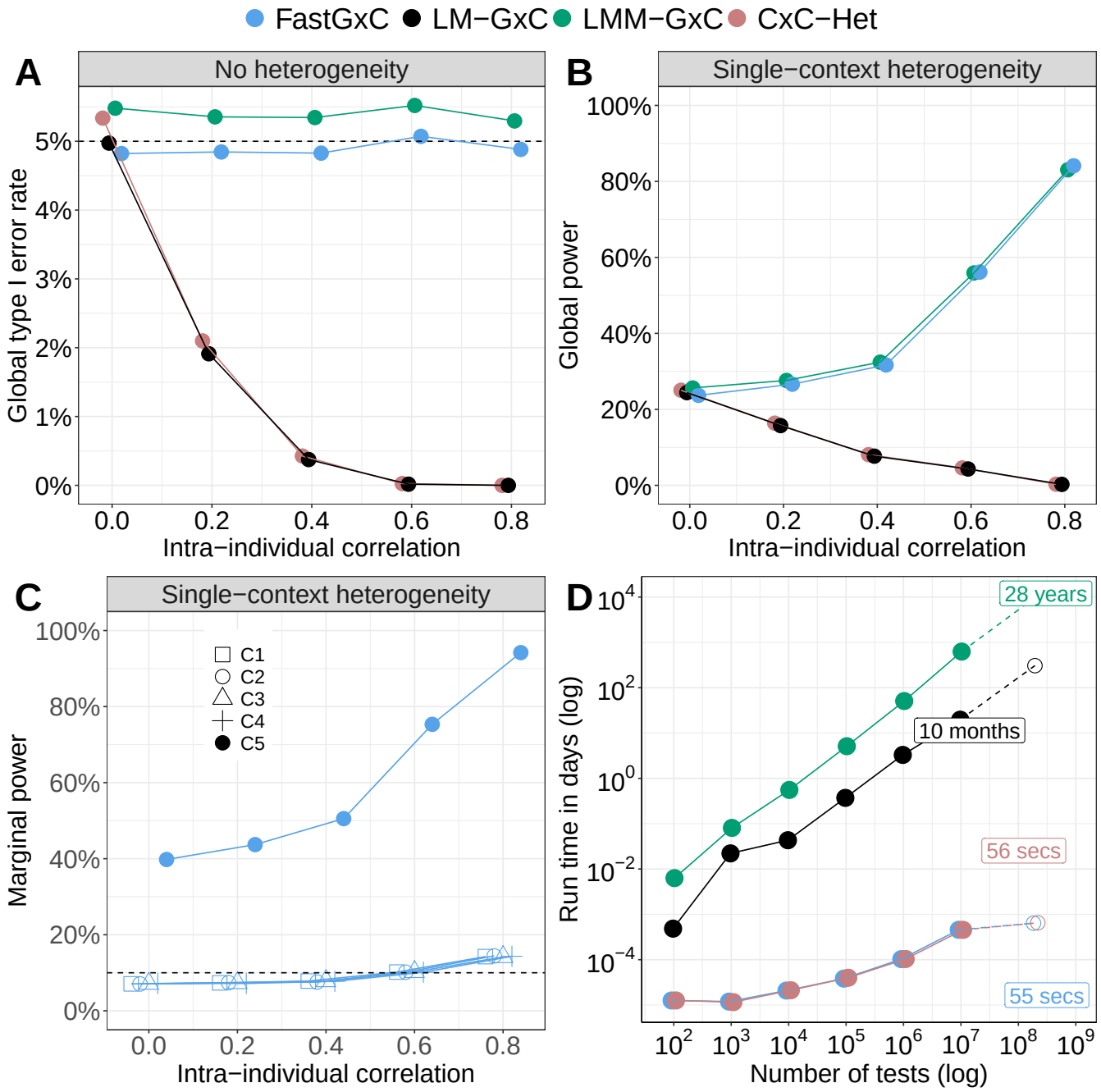


Figure 2. FastGxC outperforms existing methods in simulated data. A. Global type I error rate of all methods under different amounts of intra-individual correlation. Both LMM-GxC and FastGxC maintain proper type I error rate regardless of the intra-individual correlation while the CxC and LM-GxC approaches become more conservative with increasing amount of intra-individual correlation. **B.** Global power of all methods under the single-context heterogeneity scenario (Figure 1B). FastGxC is as powerful as the LMM-GxC approach with power increasing as a function of the amount of intra-individual correlation for both methods. The CxC and LM-GxC approaches lose power in the presence of intra-individual correlation. **C.** Marginal power of FastGxC to identify the (most) heterogeneous context under the single-context heterogeneity scenario. **D.** Run time for all methods for varying number of tests performed in a sample size of 250 individuals. See Figure S3D for sample size of 1000 individuals. Last points reflect projected run time for entire GTEx data-set - 50 contexts, 25K x 3M tests, and 250 samples per context. Analyses were run on 8 cores on a 2.70 GHz Intel Xeon Gold Processor on the UCLA Hoffman2 Computing Cluster.

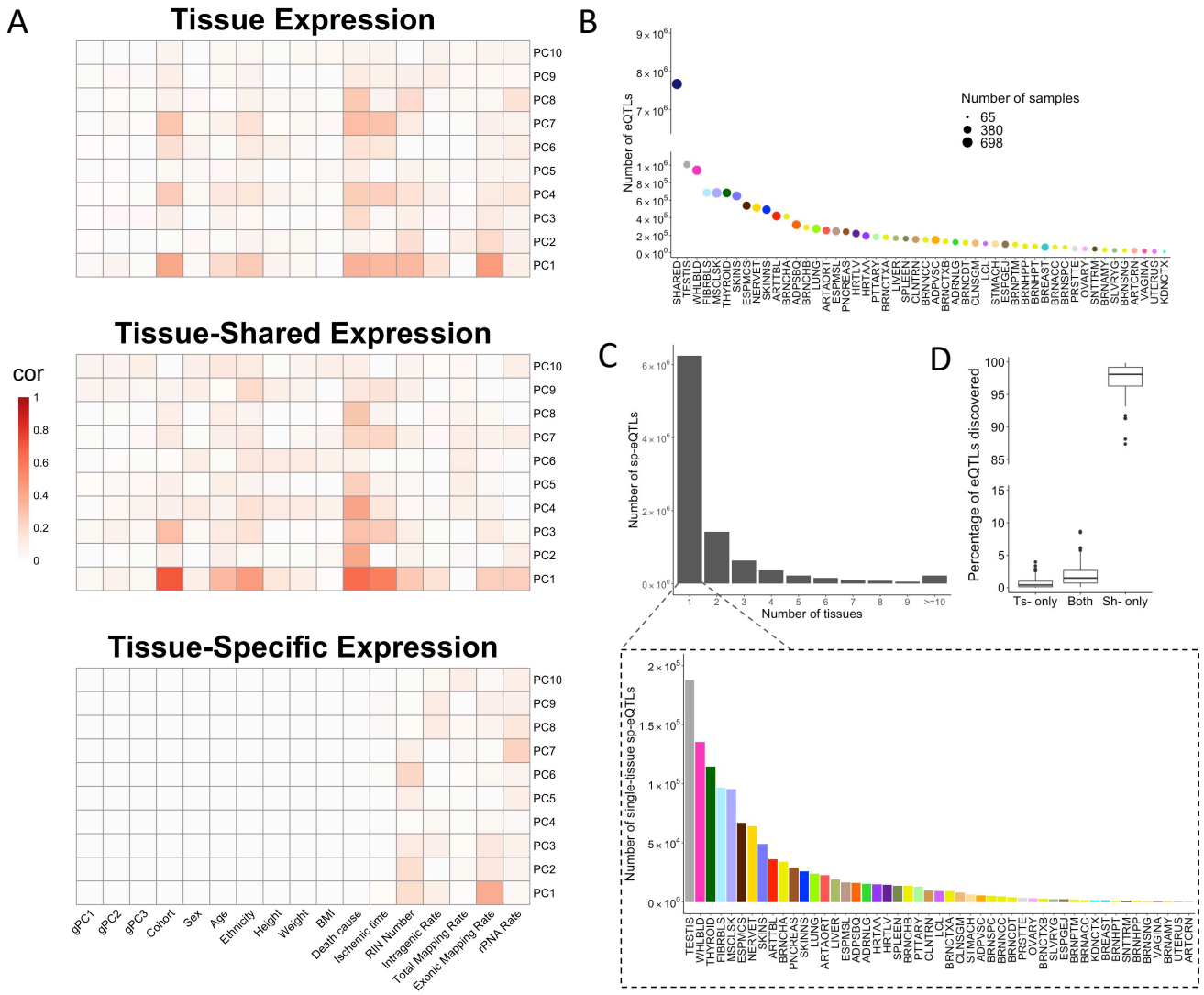


Figure 3. Tissue-specific eQTL mapping in GTEx.A. Correlation of PCs from tissue expression, tissue-shared expression, and tissue-specific expression with covariates related to study design and sample quality in GTEx. The decomposition removes the intra-individual correlation as demonstrated by lack of correlation between PCs from the tissue-specific expression and variables that are shared/invariant within an individual across tissues, e.g. genotype PCs (gPC), sex, age, etc. **B.** Number of sh-eQTLs and sp-eQTLs in each tissue. Point size reflects number of samples for each tissue. **C.** Sharing and specificity of sp-eQTLs across tissues. Top: Number of tissues with sp-eQTL effects. Bottom: Number of single tissue sp-eQTLs per tissue. **D.** Percent of eQTLs with sp-only, sh-only, and both sp- and sh- ("both") effects across all tissues. The majority of eQTLs have only shared effects and most sp-eQTLs manifest within the shared effect loci.

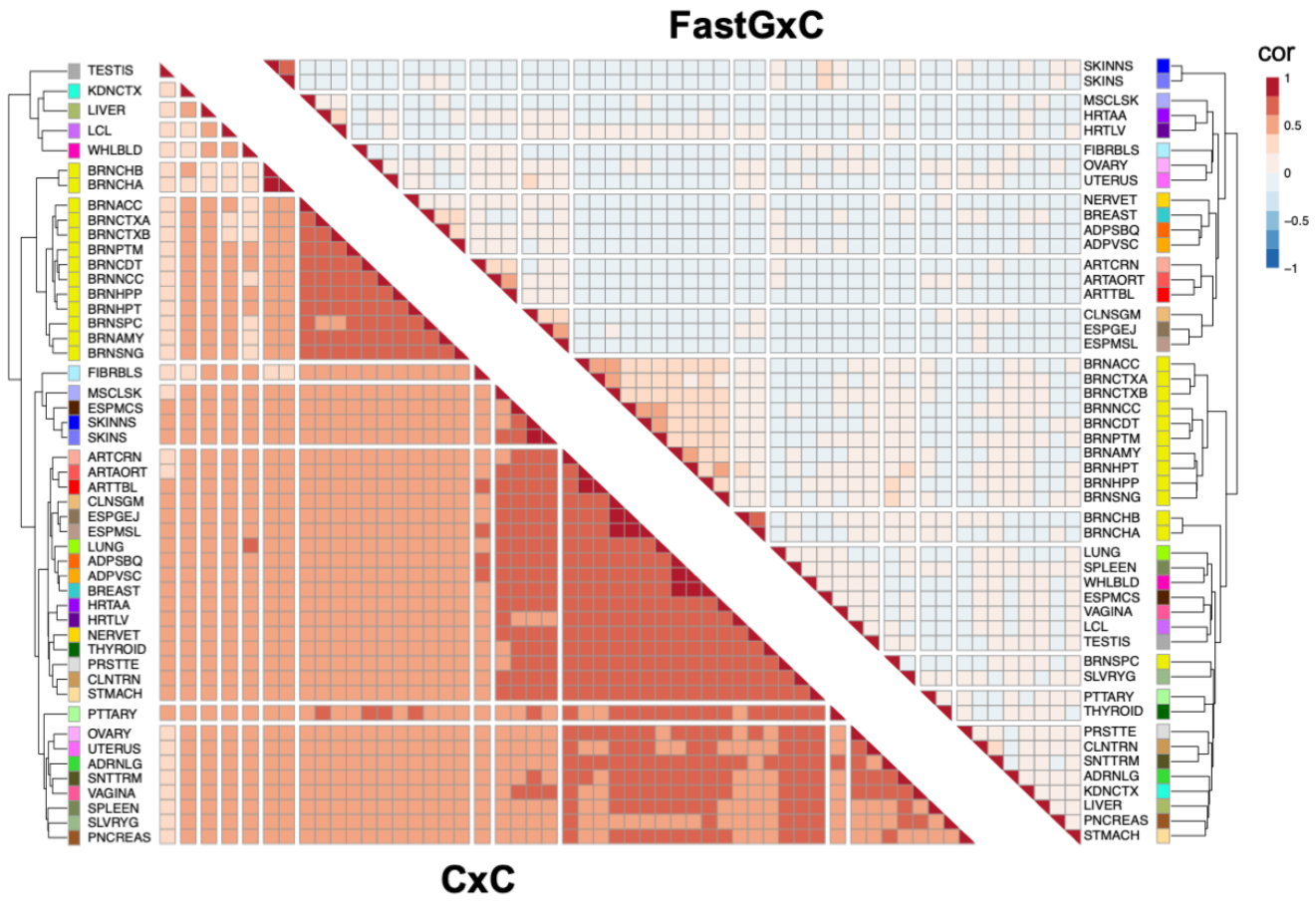


Figure 4. Tissue-specific eQTL effect sizes are correlated only between groups of closely-related tissues. Pearson correlation of eQTL effect sizes across tissues. Right: FastGxC sp-eQTL effect sizes are highly correlated only across related tissues. Largest cluster after hierarchical clustering contains brain tissues, while remaining clusters are of roughly equal size and contain tissues from related organ systems, i.e. integumentary, cardiovascular, digestive, etc. Left: CxC eQTL effect sizes are highly correlated across both groups of biologically-related and unrelated tissues. Largest cluster after hierarchical clustering on the CxC correlation matrix contains tissues from the cardiovascular and digestive systems.

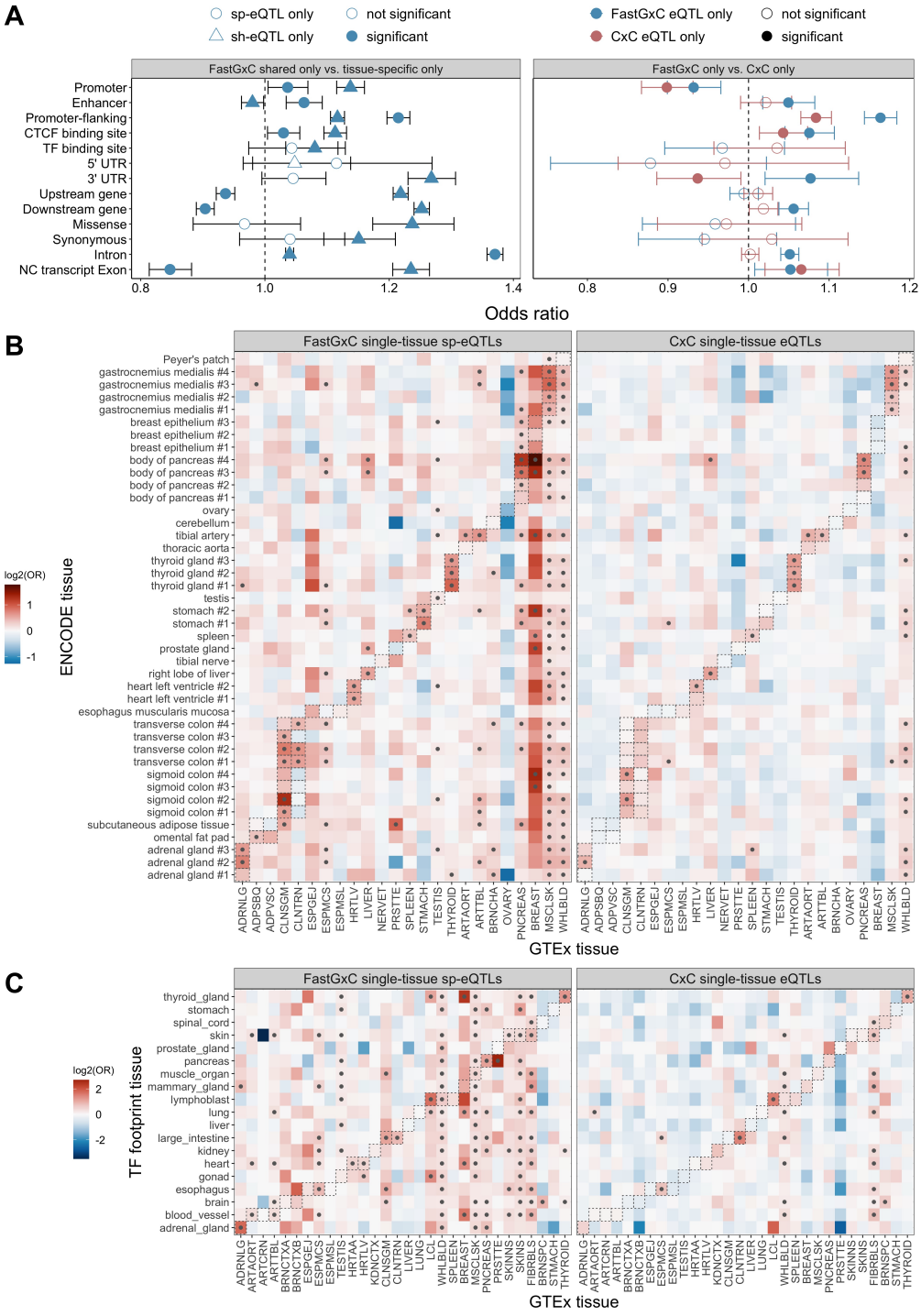


Figure 5. FastGxC sp-eQTL variants are enriched in functional genomic features from their respective tissues. **A.** Enrichment of variants with only sp-eQTL versus only sh-eQTL FastGxE effects **(left)** and only FastGxC eQTL (sh- or sp- eQTLs) versus only CxC eQTL effects **(right)** in genomic elements with known regulatory effects. **B-C.** Enrichment of single-tissue eQTLs in ATAC-seq peaks from ENCODE (B) and TF binding sites [47] (C). Boxes indicate manual matching between ENCODE ATAC-seq or TF footprint tissues and GTEx tissues. Circle indicates statistically significant enrichment. FastGxC single-tissue sp-eQTLs are more often enriched for the correct tissue and exhibit stronger enrichment than their CxC counterparts. TF: transcription factor.

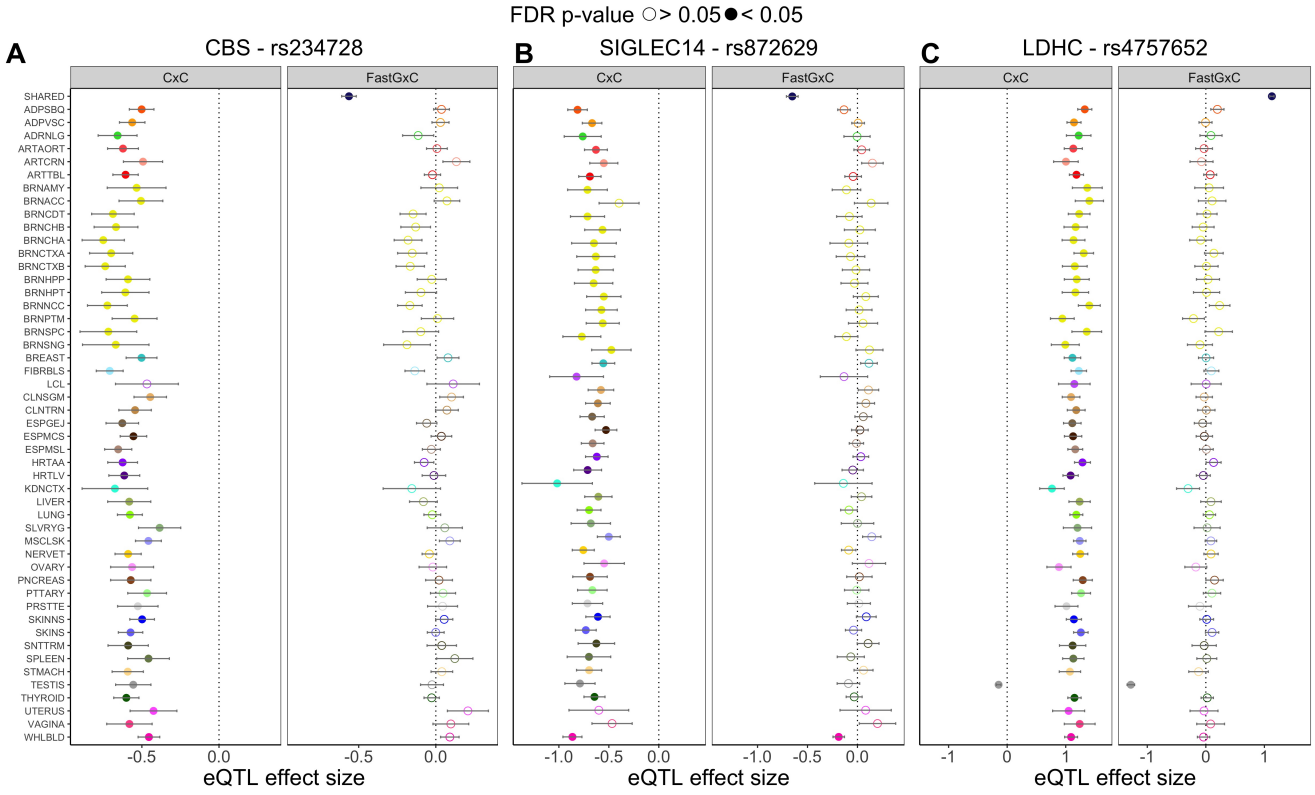
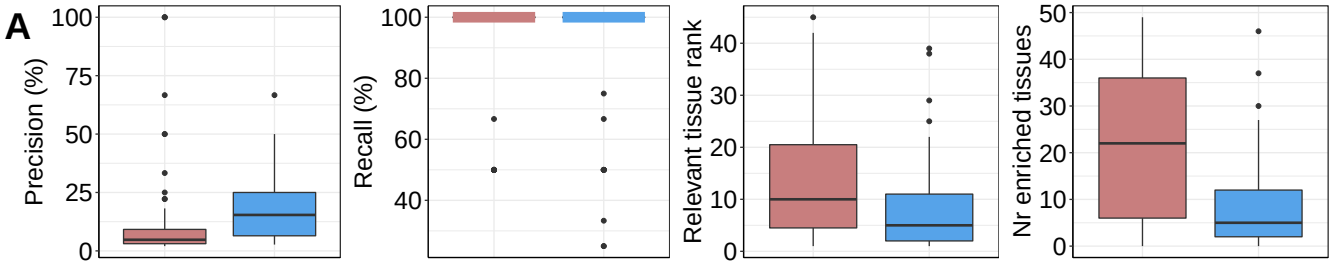


Figure 6. Examples of eQTLs identified in GTEx. Each dot shows the effect size estimates from CxC (L) and FastGxC (R) for a single tissue (color). **A.** An eQTL for the gene *CBS* shows widespread sharing across GTEx tissues captured as 48 significant CxC eQTL effects. FastGxC maps this genetic effect as a single sh-eQTL. **B.** An eQTL for the gene *SIGLEC14* shows extensive sharing across GTEx tissues captured as 47 significant CxC eQTL effects with similar effect sizes. However, after modeling the sharing as a sh-eQTL, FastGxC also maps a sp-eQTL in whole blood, consistent with the known role of *SIGLEC14* in the immune system. **C.** An eQTL for the gene *LDHC*, which acts primarily in testis, exhibits a strong positive effect in all tissues except the testis for which the eQTL effect is in the opposite direction. This lone testis-specific effect becomes very apparent when we examine sp-eQTLs with FastGxC.



B

GWAS Trait	Method	Relevant Tissue(s) Enriched (Rank)	Nr Other Tissues Enriched	Other Enriched Tissue(s)
breast carcinoma	FastGxC	BREAST (1)	16	BRNHPT, BRNCDT, BRNCTXA, BRNPTM, BRNCTXB, BRNCHA, BRNNCC, FIBRBLS, WHLBLD, BRNCHB, ADPSBQ, NERVET, THYROID, MSCLSK, SKINS, SHARED
	CxC	BREAST (25)	38	LCL, SPLEEN, BRNNCC, CLNTRN, LIVER, BRNSPC, HRTLV, CLNSGM, PRSTTE, OVARY, ADRNLG, VAGINA, UTERUS, BRNCTXB, ADPSBQ, BRNACC, ESPMCS, WHLBLD, LUNG, SLVRYG, PNCREAS, BRNHPT, BRNCDT, ESPMSL, ARTAORT, MSCLSK, SKINS, TESTIS, ADPVSC, STMACH, THYROID, PTTARY, NERVET, ESPGEJ, HRTAA, FIBRBLS, ARTTBL, SKINNS
lung adenocarcinoma	FastGxC	LUNG (2)	1	NERVET
	CxC	LUNG (18)	21	UTERUS, ARTAORT, OVARY, ADPVSC, HRTLV, SNTTRM, ESPMCS, SKINS, PNCREAS, HRTAA, ARTTBL, BRNCHA, THYROID, FIBRBLS, BREAST, SKINNS, NERVET, ADPSBQ, WHLBLD, ESPGEJ, ESPMSL
melanoma	FastGxC	SKINNS (1), SKINS (2)	2	WHLBLD, SHARED
	CxC	SKINS (1), SKINNS (5)	6	BREAST, NERVET, WHLBLD, ADPSBQ, ESPMCS, FIBRBLS
prostate carcinoma	FastGxC	PRSTTE (1)	1	SHARED
	CxC	No enrichments	0	No enrichments
testicular germ cell tumor	FastGxC	TESTIS (1)	1	SHARED
	CxC	TESTIS (10)	11	BRNCHA, PNCREAS, ESPMCS, PTTARY, FIBRBLS, ADPVSC, ARTAORT, CLNTRN, ESPMSL, MSCLSK, SKINS
cancer	FastGxC	SHARED (1)	1	WHLBLD
	CxC	No enrichments	4	WHLBLD, ADPSBQ, BREAST, ESPMCS
coronary artery disease	FastGxC	ARTCRN (1), ARTAORT (2), HRTLV (5), HRTAA (8)	13	SKINNS, ARTTBL, SHARED, MSCLSK, SPLEEN, LIVER, BRNCHA, THYROID, ESPMCS, WHLBLD, NERVET, TESTIS, FIBRBLS
	CxC	ARTAORT (6), ARTCRN (11), HRTAA (23), HRTLV (32)	39	STMACH, CLNTRN, ARTTBL, ADPSBQ, ESPMSL, BRNACC, NERVET, SKINNS, LUNG, FIBRBLS, BREAST, OVARY, UTERUS, LIVER, BRNHPP, PNCREAS, SNTTRM, VAGINA, ESPMCS, SKINS, MSCLSK, SLVRYG, LCL, BRNCTXA, BRNCHB, CLNSGM, PRSTTE, WHLBLD, BRNNCC, ADPVSC, THYROID, BRNCHA, ADRNLG, BRNHPT, PTTARY, BRNPTM, TESTIS, ESPGEJ, SPLEEN
atrial fibrillation	FastGxC	HRTAA (1), HRTLV (2)	4	PTTARY, FIBRBLS, WHLBLD, SHARED
	CxC	HRTLV (1), HRTAA (2)	11	BRNHPP, PRSTTE, ESPMSL, ESPGEJ, THYROID, PNCREAS, BREAST, TESTIS, MSCLSK, ADPVSC, ARTTBL

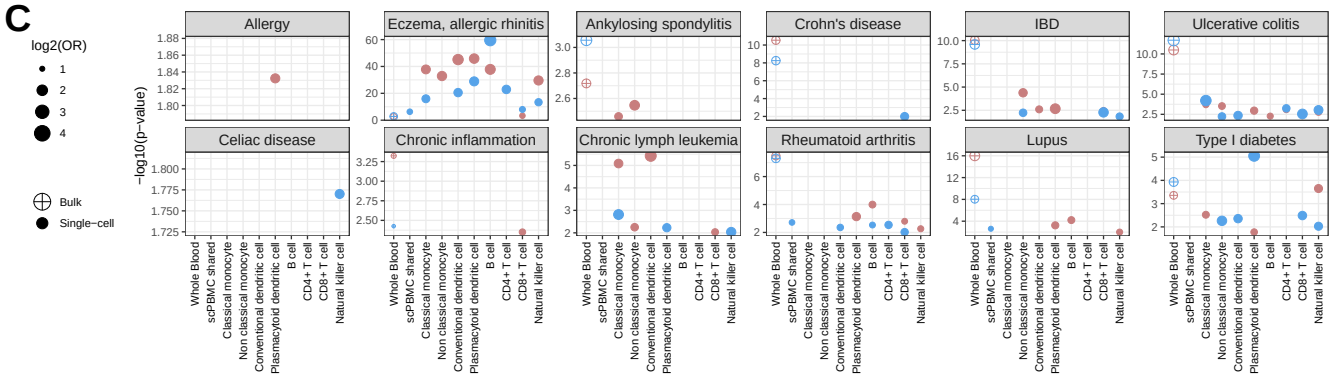


Figure 7. FastGxC identifies context-relevant mechanisms of complex traits. A. Accuracy of FastGxC and CxC eQTLs to prioritize the most relevant tissue(s) across 138 complex traits with a strong prior indication for the likely relevant tissue(s). **B.** Tissues prioritized by FastGxC and CxC as well as the rank of the known relevant tissues for specific complex traits. **C.** Enrichment of FastGxC shared and cell-type-specific eQTLs and CxC eQTLs mapped in each cell type (x-axis) for a set of trait-associated variants from 59 immune-related traits in the NHGRI-EBI GWAS catalog. For comparison, we include enrichment results from GTEx CxC whole blood eQTLs and FastGxC whole-blood-specific eQTLs. Each dot represents the enrichment p-value and the size represents the log2 odds ratio (OR) of a significant cell type-trait enrichment.

Online Methods

Relationship between FatsGxC, CxC, and LM(M)-GxC parameters. Let β_c be the CxC eQTL effect in context c, as estimated by fitting a linear regression model per context, i.e., $E_{ic} = \alpha_c + \beta_c G_i + \varepsilon_{ic}$. Then, the CxC eQTL effect in context c is equal to the sum of the shared and context-c-specific eQTL effects from FastGxC, i.e. $\beta_c = \beta^{sh} + \beta_c^{cs}$. In addition, let β_{ref} be the eQTL effect in an arbitrarily defined reference tissue and δ_c be the interaction eQTL effects for the non-reference tissues c from an L(M)M model with a genotype-by-context interaction term, i.e. $E_{ic} = (u_i) + \alpha + \beta_1 G_i + \sum_{c=2}^{C} \gamma_c K_{ic} + \sum_{c=2}^{C} \delta_c G_i \times K_{ic} + \varepsilon_{ic}$. Then, $\beta_{ref} = \beta^{sh} + \beta_{ref}^{cs}$ and $\delta_c = \beta_c - \beta_{ref} = \beta^{sh} + \beta_c^{cs} - \beta^{sh} - \beta_{ref}^{cs} = \beta_c^{cs} - \beta_{ref}^{cs}$ for $c \neq ref$. Full details of the analytical derivation are provided in the Supplementary Text.

GTEx data. Fully processed, filtered, and normalized gene expression matrices (in BED format) for each tissue as well as covariates which were used as input for eQTL analysis were downloaded through the GTEx portal (`https://www.gtexportal.org/home/datasets`) on March 11, 2020. Gene expression matrices were residualized for covariates. WGS genotype VCF data were downloaded from dbGap (dbGaP Accession phs000424.v8.p2). Only individuals with both genotype and gene expression data were kept. VCF files were processed with vcftools (v0.1.16) to keep only bi-allelic SNPs. Only variants with minor allele frequencies of greater than five percent in the tissue of interest were kept. Bcftools (v1.12) was used to annotate the genotype files with rs IDs. Plink (v1.90) was used to transpose and convert the vcf files to a sample x genotype matrix which was used as input for eQTL mapping.

FastGxC and CxC eQTL mapping in GTEx and CLUES. Expression of each gene was centered to have mean zero across all individuals and tissues and decomposed into 49 tissue-specific expression components and one shared expression component using FastGxC. Cis genetic effects on the shared gene expression levels, each tissue-specific gene expression levels (FastGxC), and gene expression levels in each tissue (CxC) were estimated using ultra-fast implementations of simple linear regression models as implemented in the `MatrixEQTL` R package [31] with `model=modelLINEAR`

and 1e6 basepair distance for calling cis-eQTLs. Multiple testing correction was performed using the hierarchical FDR procedures implemented in the R package `TreeQTL` [58] with genes in level one, genes-tissues in level two, and genes-tissues-SNPs in level three. eQTL mapping in the single-cell CLUES data is described in detail here (see accompanying manuscript). Multiple testing correction was performed using hierarchical FDR with genes in level one, genes-cell-types in level two, and genes-cell-types-SNPs in level three.

Correlation between PCs and covariates in GTEx The correlation between expression PCs and covariates in GTEx was computed using the `canCorPairs` function from the `variancePartition` R package ([59]). In short, when comparing two continuous variables (e.g. gPC1 or weight), Pearson correlation was used. In order to accommodate the correlation between a continuous and a categorical variable (e.g. cohort) canonical correlation analysis (CCA) was used. Note that CCA returns correlations values between 0 and 1.

Background SNP-gene pairs for enrichment analyses For all enrichment analysis, the `matchit` function from the `MatchIt` R package was used to match a set of background SNP-gene pairs to each variant set of interest by minor allele frequency (MAF) using the nearest neighbor matching method and a 1:1 matching ratio [60]. For eQTL sets that contained more than 5000 variants, sets were randomly split into chunks to speed up computation.

EQTL enrichment in genomic features Sp-eQTL only and sh-eQTL only variant sets were obtained by taking the set difference of sp-eQTL and sh-eQTL variants in R, respectively. The FastGxC eQTL variant set was obtained by taking the union of sh- and sp-eQTL variants across tissues. The CxC eQTL variant set was obtained by taking the union of eQTL variants across tissues. The set difference of FastGxC eQTL variant set and the CxC eQTL variants were then computed by taking the set difference in R to obtain the final FastGxC-only and CxC-only eQTL variant sets. All variants that were used as input into MatrixEQTL were inputted into the Ensembl Variant Effect Predictor (VEP) tool, which determines the effects of variants such as consequence on protein sequence or location within genomic regulatory elements. Enrichment analysis was then performed using the EQTL sets as described above and the VEP annotated variants list by

performing a Fisher's exact test from the R `stats` package followed by a Benjamini and Hochberg multiple testing adjustment. Significance was called for BH-adjusted p-values less than 0.05.

Enrichment in ENCODE ATAC-seq data All available tissue ATAC-seq data in the "not perturbed", GRCh38, and bigBed narrowPeak categories were downloaded from `www.encodeproject.org` on Novemeber 2020. The downloaded bigBed files were converted to bed files for downstream analysis by the UCSC bigbedtobed tool. Bed files were then sorted using the `bedtools sort -k1,1 -k2,2n` command to enable a memory-efficient algorithm for downstream intersections. Enrichment analysis of FastGxC and CxC single-tissue eQTL variants was then performed by intersecting each eQTL variant set of interest with each pre-sorted bed file, corresponding to ATAC-seq peaks from one tissue/sample, using the bedtools intersectBed command. Finally, Fisher's exact test was used to obtain the statistical significance of each enrichment, followed by a Benjamini and Hochberg multiple testing adjustment. Significance was called for BH-adjusted p values less than 0.05.

Enrichment in Transcription factor binding sites Transcription factor binding site data was downloaded on October 2020 from `http://data.nemoarchive.org/other/grant/sament/sament/footprint_atlas/bed/` using the HINT algorithm and 16 basepair seed length. To constrain analysis to the top footprints, the data was filtered using a HINT score greater than 200, as described by the method authors as an ideal threshold for high quality footprints [47]. TF footprint genomic intervals were sorted using the bedtools sort command as described above. Finally, enrichment of eQTL variant sets were performed by intersecting variants with TF footprints of each tissue using the bedtools intersectBed command. Fisher's exact test was used to obtain the statistical significance of each enrichment, followed by a Benjamini and Hochberg multiple testing adjustment. Significance was called for BH-adjusted p values less than 0.05.

Enrichment in GWAS loci Genome-wide association study (GWAS) data (gwas_catalog_v1.0.2-associations_e100_r2020-06-17) was downloaded and processed from the NHGRI-EBI GWAS Catalog in August 2020 [53]. Matching of variants with and without eQTL effects was performed as described above. Only mapped traits within the GWAS catalog that contained more than ten variants were included in our downstream workflow. Enrichment analysis of FastGxC and CxC eQTL variants

was then performed by intersecting each eQTL variant set of interest with variants from each mapped trait by rs ID. Finally, Fisher's exact test was used to obtain the statistical significance of each enrichment. A hierarchical multiple testing procedure was performed by first obtaining Simes's method for combining p-value per tissue across mapped traits, BH-adjusting the resulting 49 tissue-level p-values, and then retaining only tissues with BH-adjusted Simes' p-values under the tissue-level α of 0.05. Then, within each significant tissue, p-values across all mapped traits were BH-adjusted and filtered using a trait-level α, i.e. tissue-level α * (n_significant_tissues / n_total_tissues) to obtain the final significant tissue-trait associations.

Precision and recall of context-relevant mechanisms of complex traits. We manually annotated 138 traits within the GWAS Catalog with their most likely tissue of interest and used this annotation to assign precision and recall rates. More specifically, we used a contingency table, per trait, by calculating how often the trait of interest is both enriched in a tissue's eQTLs and the tissue is the assigned likely-relevant tissue, giving true/false positive/negative rates (TP, FP, TN, FN). Finally, the precision score was calculated as TP / (TP + FP), and the recall score was calculated as TP / (TP + FN).

Supplementary Material

Exact relation between FastGxC and CxC estimates

Fix a gene and assume that its expression in each context follows a linear model:

$$E^0_{i,} = G\beta^0 + \epsilon_{i,} \in \mathbb{R}^C$$

where:

- $E^0 \in \mathbb{R}^{N\times C}$ is the matrix of gene expression across N samples and C contexts, e.g. tissues or cell types
- $G \in \mathbb{R}^{N\times S}$ is an arbitrary covariate matrix containing S features (in this paper, the features are *cis*-SNPs, and usually $S = 1$)
- $\beta^0 \in \mathbb{R}^{S\times C}$ are the context-specific effects captures arbitrarily distributed noise, assumed i.i.d. over samples i
- $\epsilon_{i,} \in \mathbb{R}^C$ captures arbitrarily distributed noise, assumed i.i.d. over samples i but with covariance between contexts given by $\mathbb{V}(\epsilon_{i,}) = \Sigma$

Now define the context-centered expression as:

$$E_i = E^0_i - \bar{E}_i 1^T_C \quad \text{or} \quad E_{ic} = E^0_{ic} - \bar{E}_i$$

where $1_C \in \mathbb{R}^C$ is a vector of 1s and $\bar{E} \in \mathbb{R}^N$ is a vector containing each sample's mean expression across all C contexts.

For any arbitrary vector $X \in R^{1\times N}$, we have:

$$XE = XE^0 - X\bar{E}1^T_C$$

In particular, when $X = \frac{1}{\|G_j\|^2}G_j$ for SNP j, then:

- $XE := \hat{\beta}$ gives the FastGxC cs-eQTL effect size estimates for SNP j
- $XE^0 := \hat{\beta}^0$ gives the ordinary cs-eQTL effects

- $X\bar{E} := \bar{\beta}$ gives the FastGxC sh-eQTL effects

Putting these three facts together proves:

$$\hat{\beta}_c = \hat{\beta}_c^0 - \bar{\beta} \quad \text{or} \quad \text{FastGxC} = \text{CxC} - \text{Shared}$$

for all contexts c. In words, the standard contest-specific estimates in CxC naturally and exactly decouple into the FastGxC estimates and the cross-context average estimate.

By the same argument, CxC decomposes into FastGxC and shared effects even when:

- Covariates are included, via $X = \frac{1}{\|P_Z^\perp G_j\|^2} G_j P_Z^\perp$, where $P_Z^\perp$ is the orthogonal projection onto the span of the covariate matrix Z
- Multiple SNP effects are fit simultaneously, via $X = (GG^T)^{-1}G^T$
- Ridge regression/kinship-based LMMs are used, if the regularization/heritability is equal across contexts

Conceptually, associativity guarantees that linear operators applied to the left of the matrix E play well with linear operators applied to its right. And most regression involve linear operations on E from the left, while the centering operation used by GxC is a linear operator from the right. That is, we can center and then perform regressions (as in FastGxC) or can perform regular regressions and then center; these operations associate, therefore give identical results.

Approximate relation between FastGxC and CxC standard errors

Above, we showed the CxC estimates exactly decouple into FastGxC and shared estimates. Here, we show a similar result for the standard errors, though it holds only approximately. Specifically, the variance of the FastGxC estimate is roughly the variance of the CxC estimate minus the variance in the shared estimate. This provides a sharp description of the improvement in power in FastGxC over CxC due to removal of shared noise.

More concretely, using the equivalence proved above, we have:

$$\begin{aligned}
\mathbb{V}\left(\hat{\beta}_c\right) &= \mathbb{V}\left(\hat{\beta}_c^0 - \bar{\beta}\right) \\
&= \mathbb{V}\left(\hat{\beta}_c^0\right) + \mathbb{V}\left(\bar{\beta}\right) - 2\mathrm{Cov}(\hat{\beta}_c^0, \bar{\beta}) \\
&= \mathbb{V}\left(\hat{\beta}_c^0\right) - \mathbb{V}\left(\bar{\beta}\right) - 2\left(\mathrm{Cov}(\hat{\beta}_c^0, \bar{\beta}) - \mathbb{V}\left(\bar{\beta}\right)\right) \\
&\approx \mathbb{V}\left(\hat{\beta}_c^0\right) - \mathbb{V}\left(\bar{\beta}\right) \qquad (*)
\end{aligned}$$

Loosely, the approximation assumes that the contexts are roughly exchangeable, or that each context is roughly equally correlated with other contexts[1]. For example, this holds exactly in the cases where contexts are IID ($\Sigma = \sigma^2 I$) or exchangeable ($\Sigma = \sigma^2 I + bJ$); conversely, this is violated if context c is very unique, or if there large and structured subsets of the contexts (eg brain regions).

For example, imagine that C is large and that each sample's noise has exchangeable distribution across contexts, implying that $\mathbb{V}(\epsilon_{i,}) = \sigma^2 I + sJ$ for some $\sigma^2 > c$. Then the above approximation is exact, and standard error in FastGxC simply subtracts off the standard error for the shared noise term, s:

$$\mathbb{V}\left(\hat{\beta}_c\right) = \frac{1}{\|X\|^2}(\sigma^2 + s) - \frac{1}{\|X\|^2}(\frac{1}{C}\sigma^2 + s) \approx \frac{1}{\|X\|^2}\sigma^2$$

[1]More formally, if we assume that ϵ_i are i.i.d. with cross-context covariance matrix Σ, then:

$$\mathrm{Cov}(\hat{\beta}_c^0, \bar{\beta}) = \mathrm{Cov}(XE_{,c}^0, X\bar{E}) = X\mathrm{Cov}(E_{,c}^0, \bar{E})X^T = \|X\|^2\mathrm{Cov}(E_{,c}^0, \frac{1}{C}E^0 1_C) = \frac{1}{C}\|X\|^2\Sigma_{c,}1_C = \|X\|^2\Sigma_{c.}$$

$$\Sigma_{c.} := \frac{1}{C}\sum_{c'}\Sigma_{cc'}$$

and likewise (using $\otimes$ for tensor/Kronecker product, and $\mathrm{vec}(\cdot)$ for column-wise matrix vectorization):

$$\mathbb{V}\left(\bar{\beta}\right) = \mathbb{V}\left(X\bar{E}\right) = \mathbb{V}\left(\left((\frac{1}{C}1_C^T) \otimes X\right)\mathrm{vec}(E^0)\right) = \frac{1}{C^2}\left(1_C^T \otimes X\right)\left(\Sigma \otimes I_N\right)\left(1_C^T \otimes X\right)^T = \left(\frac{1}{C^2}1_C^T\Sigma 1_C\right)\cdot(XX^T) = \|X\|^2\Sigma_{..}$$

$$\Sigma_{..} := \frac{1}{C^2}\sum_{c,c'}\Sigma_{cc'}$$

Thus, $(*)$ assumes that $\Sigma_{c.} \approx \Sigma_{..}$, i.e. that context c is about as correlated with the average context as any other.

Inter-context noise correlation does not affect FastGxC estimates

Say that samples are i.i.d. Gaussian but that contexts are correlated:

$$E_{i,} \overset{\text{iid}}{\sim} G_{i,}\beta^0 + \mathcal{N}(0, \Sigma)$$

Assume that we estimated or know the noise covariance Σ, e.g. with an LMM. The GLS and OLS estimates for B are identical–again, conceptually, the key fact is that column transformations on E operate independently of row transformations. (Σ acts on the rows of E, while G acts on the columns.) One way to see this is using the covariance across all entries of E, $\mathbb{V}(\text{vec}(E)) = \Sigma \otimes I_N$:

$$\begin{aligned}
\hat{\beta}_{GLS} &:= \left((G \otimes I_P)^T (I_N \otimes \Sigma)^{-1} (G \otimes I_P)\right)^{-1} (G \otimes I_P)^T (I_N \otimes \Sigma)^{-1} \text{vec}(E) \\
&= \left((G^T G)^{-1} \otimes \Sigma\right) \left(G^T \otimes \Sigma^{-1}\right) \text{vec}(E) \\
&= \left(((G^T G)^{-1} G^T) \otimes I_P\right) \text{vec}(E) \\
&= \text{vec}\left((G^T G)^{-1} G^T E\right) \\
&= \hat{\beta}_{OLS}
\end{aligned}$$

tissue	color_hex	abbreviation
Adipose_Subcutaneous	#FF6600	ADPSBQ
Adipose_Visceral_Omentum	#FFAA00	ADPVSC
Adrenal_Gland	#33DD33	ADRNLG
Artery_Aorta	#FF5555	ARTAORT
Artery_Coronary	#FFAA99	ARTCRN
Artery_Tibial	#FF0000	ARTTBL
Brain_Amygdala	#EEEE00	BRNAMY
Brain_Anterior_cingulate_cortex_BA24	#EEEE00	BRNACC
Brain_Caudate_basal_ganglia	#EEEE00	BRNCDT
Brain_Cerebellar_Hemisphere	#EEEE00	BRNCHB
Brain_Cerebellum	#EEEE00	BRNCHA
Brain_Cortex	#EEEE00	BRNCTXA
Brain_Frontal_Cortex_BA9	#EEEE00	BRNCTXB
Brain_Hippocampus	#EEEE00	BRNHPP
Brain_Hypothalamus	#EEEE00	BRNHPT
Brain_Nucleus_accumbens_basal_ganglia	#EEEE00	BRNNCC
Brain_Putamen_basal_ganglia	#EEEE00	BRNPTM
Brain_Spinal_cord_cervical_c-1	#EEEE00	BRNSPC
Brain_Substantia_nigra	#EEEE00	BRNSNG
Breast_Mammary_Tissue	#33CCCC	BREAST
Cells_Cultured_fibroblasts	#AAEEFF	FIBRBLS
Cells_EBV-transformed_lymphocytes	#CC66FF	LCL
Colon_Sigmoid	#EEBB77	CLNSGM
Colon_Transverse	#CC9955	CLNTRN
Esophagus_Gastroesophageal_Junction	#8B7355	ESPGEJ
Esophagus_Mucosa	#552200	ESPMCS
Esophagus_Muscularis	#BB9988	ESPMSL
Heart_Atrial_Appendage	#9900FF	HRTAA
Heart_Left_Ventricle	#330099	HRTLV
Kidney_Cortex	#22FFDD	KDNCTX
Liver	#AABB66	LIVER
Lung	#99FF00	LUNG
Minor_Salivary_Gland	#99BB88	SLVRYG
Muscle_Skeletal	#AAAAFF	MSCLSK
Nerve_Tibial	#FFD700	NERVET
Ovary	#FFAAFF	OVARY
Pancreas	#995522	PNCREAS
Pituitary	#AAFF99	PTTARY
Prostate	#DDDDDD	PRSTTE
Skin_Not_Sun_Exposed_Suprapubic	#0000FF	SKINNS
Skin_Sun_Exposed_Lower_leg	#7777FF	SKINS
Small_Intestine_Terminal_Ileum	#555522	SNTTRM
Spleen	#778855	SPLEEN
Stomach	#FFDD99	STMACH
Testis	#AAAAAA	TESTIS
Thyroid	#005500	THYROID
Uterus	#FF66FF	UTERUS
Vagina	#FF5599	VAGINA
Whole_Blood	#FF00BB	WHLBLD
Shared	midnightblue	SHARED

Figure S1. Colors and abbreviations for GTEx tissues.

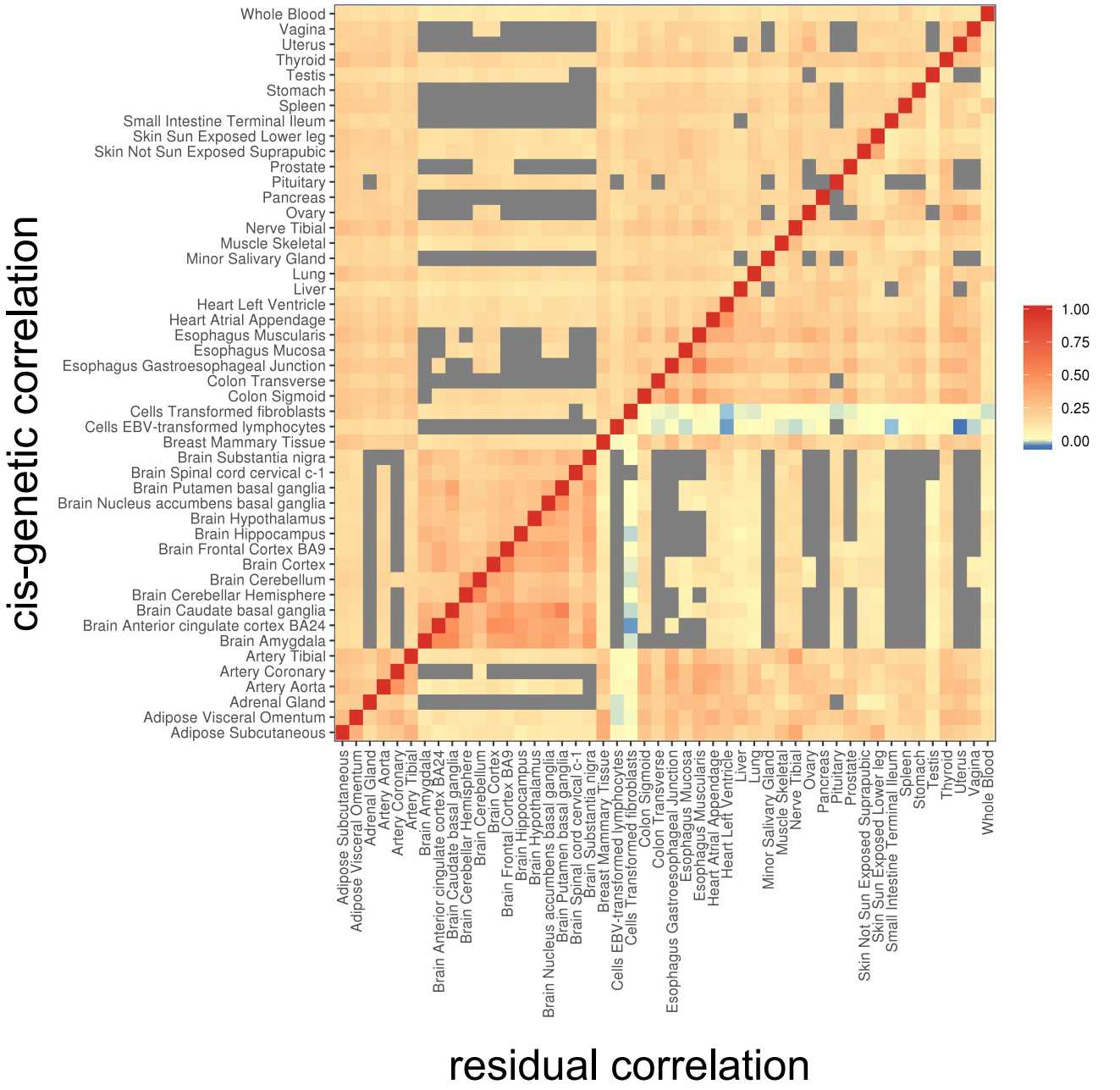


Figure S2. Genetic correlation of gene expression across tissues in the GTEx study. Cis-genetic and residual variance and covariance components for each gene-tissue pair across GTEx as calculated using a linear mixed model with bivariate REML[61]. The gray units indicate tissue pairs with less than 10% sample overlap. In both the genetic (upper) and residual (lower) components, there was widespread correlation, and the brain tissues were relatively highly correlated compared to other tissues.

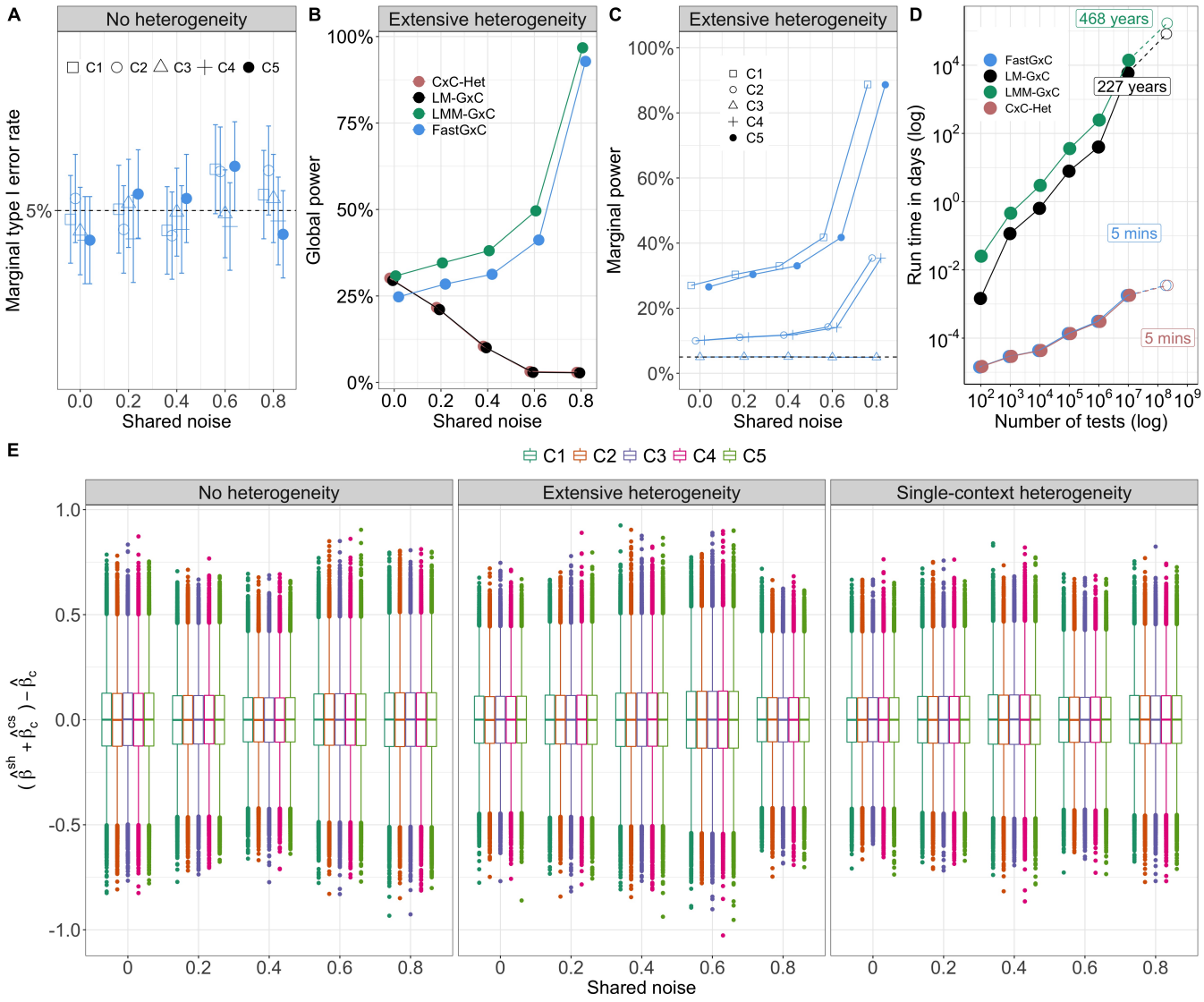


Figure S3. FastGxC performance in simulated data. **(A)** Marginal type I error rate for FastGxC under different amounts of intra-individual correlation. FastGxC maintains proper type I error rate for each context and different amounts of intra-individual correlation. **(B)** Global power of each method to identify eQTL heterogeneity under the extensive heterogeneity scenario. **(C)** Marginal power of FastGxC to identify the tissue(s) driving the eQTL effect size heterogeneity under the extensive heterogeneity scenario. **(D)** Run time of each method in a simulated scenario with 1000 individuals. **(E)** Ability of FastGxC estimates under the null and two alternative scenarios to estimate eQTL effects in each context.

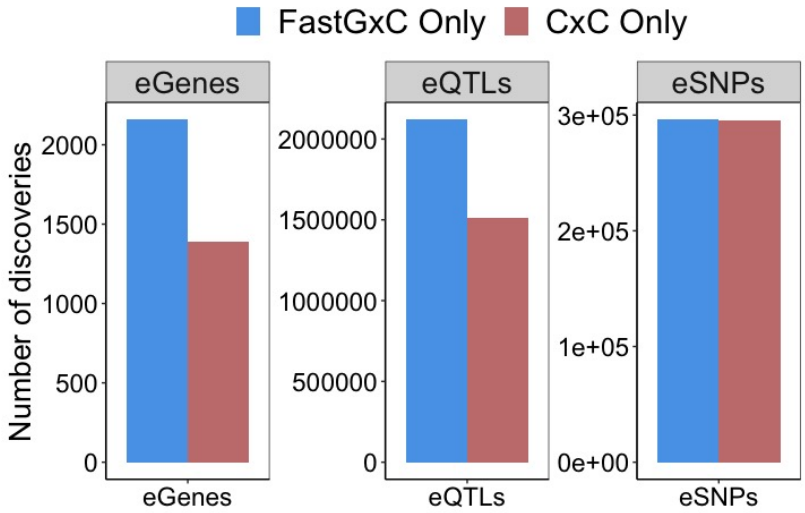


Figure S4. Comparison of FastGxC-only and CxC-only discoveries in GTEx. Comparing discoveries that are mapped uniquely by each method, FastGxC discovers more eGenes, i.e. genes with at least one sh- or sp-eQTL effects in at least one tissue, and eQTLs, i.e. gene-snp pairs with sh- or sp-eQTL effects in at least one tissue, than CxC. FastGxC and CxC map roughly the same number of eSNPs, i.e. variants with (sh- or sp-) eQTL effects in at least one tissue.

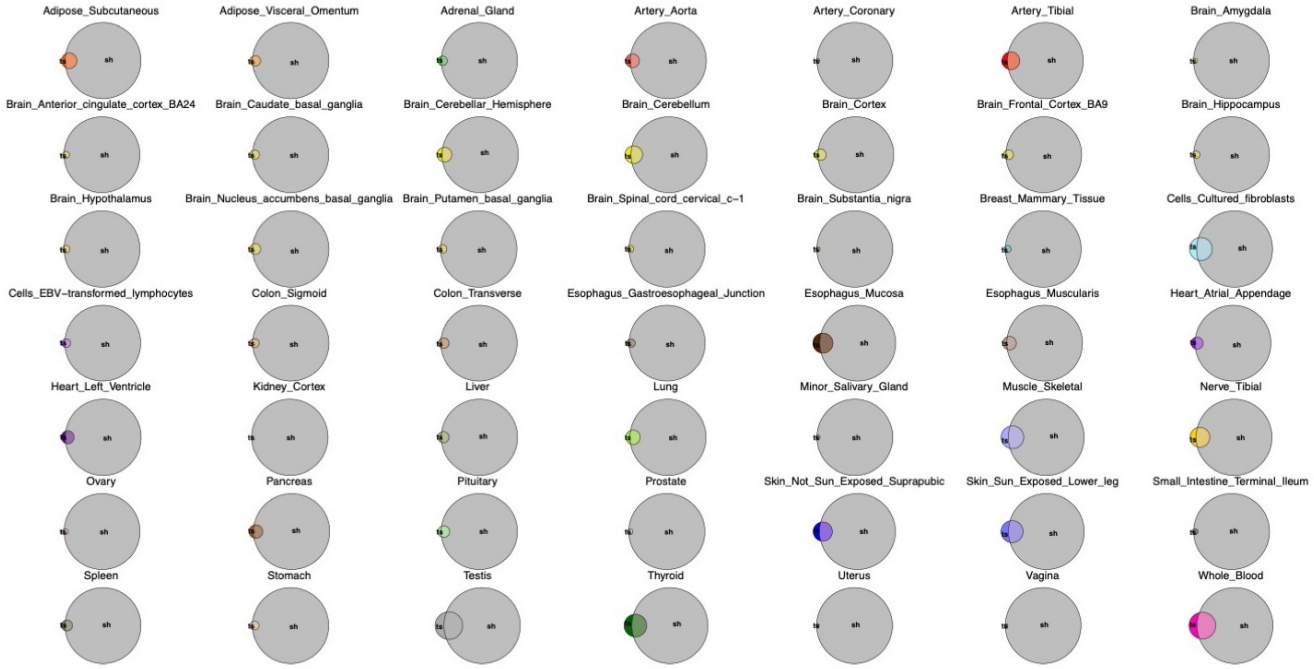


Figure S5. Comparison of FastGxC sh- and sp- eQTLs. For each tissue, we plotted Venn diagrams comparing the set of sp-eQTLs to sh-eQTLs. In the vast majority of tissues, sp-eQTLs also have sh-eQTL effects. The distribution of sharing can be found in Figure 3D.

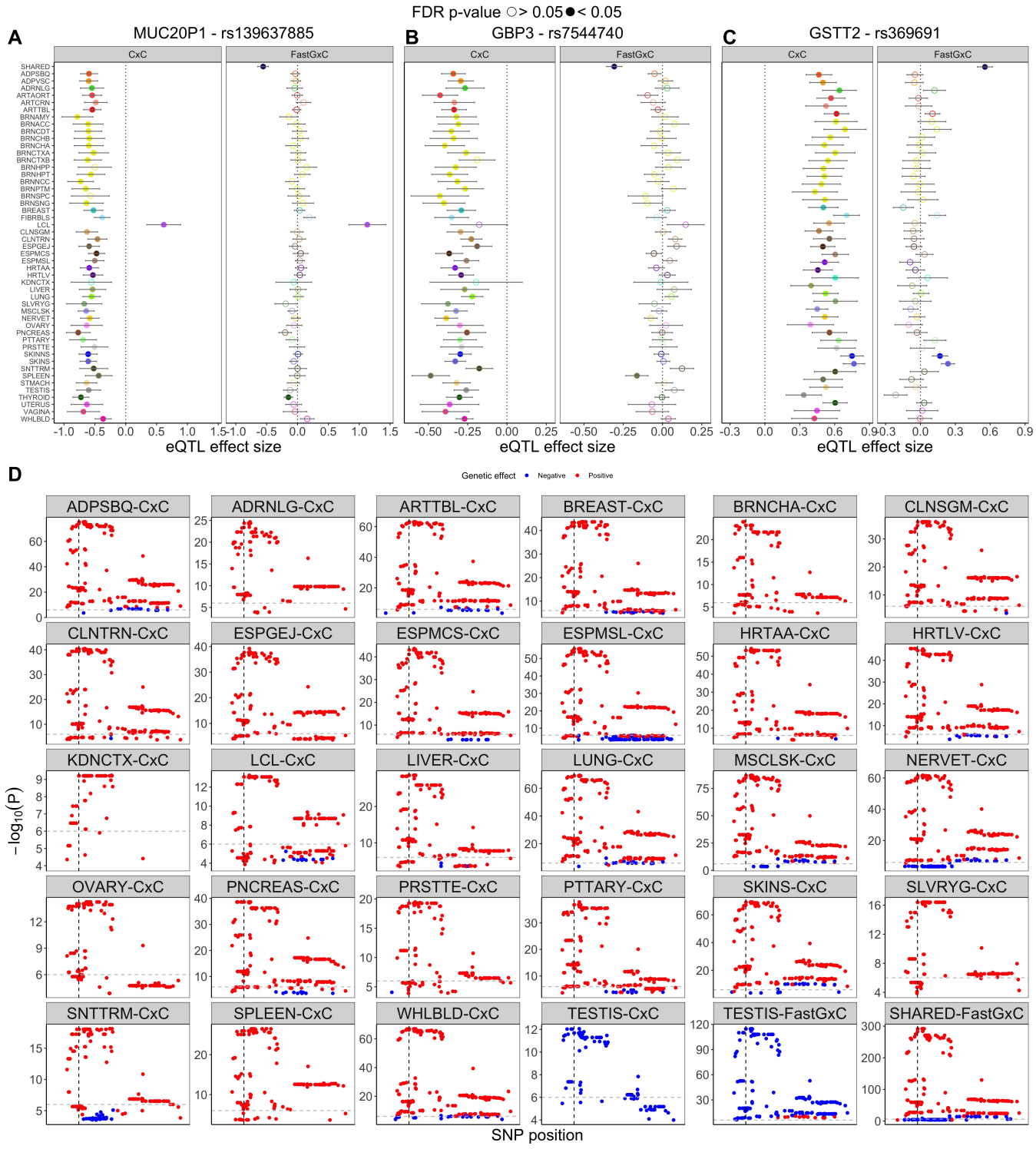


Figure S6. Examples of eQTL mapped by FastGxC in GTEx. A. An eQTL for the gene MUC20P1 and variant rs139637885 shows widespread sharing across all tissues except for LCLs. This eQTL is captured by FastGxC as a sh-eQTL, as well as a sp-eQTL in LCLs and a less-obvious sp-eQTL in thyroid. **B.** An eQTL for the gene *GBP3*. Another example of widespread sharing that obscures a tissue-specific effect in the spleen. **C.** An eQTL for the gene *GSTT2* and variant rs369691 shows widespread sharing across all tissues in CxC. After the FastGxC procedure, two sp-eQTLs emerge in the two skin tissues. **D.** Manhattan plot for LDHC eQTLs reveal a set of variants with similar genetic effects as rs4757652. These eQTLs exhibit positive CxC genetic effects across many tissues besides the testis, while the FastGxE procedure crystallizes this testis tissue specificity.

Table S1. FastGxC mapped GTEx sp-eGenes. FastGxC and CxC mapped sh- and sp-eGenes from GTEx and CLUES cohorts are provided as a separate excel file, one sheet per study, with the following columns: eGene type (CxC eGene, FastGxC sh-eGene, or FastGxC sp-eGene), tissue or cell type, gene identifier.

Table S2. EQTL enrichment in GWAS loci results. Results from enrichment of GTEx and single-cell eQTLs from CxC and FastGxC in GWAS catalog loci are provided as a separate excel file with two sheets. The first sheet shows manual annotation of most likely relevant tissue(s) for GWAS catalog traits with the following columns: GWAS Trait, Most Likely Relevant Tissue(s). The second sheet shows enrichment results with the following columns: GWAS trait, method, tissue (GTEx) or cell type (single-cell), enrichment odds ratio (OR), OR lower confidence interval, OR upper confidence interval, enrichment p-value.

References

[1] Matthew T. Maurano, Richard Humbert, et al. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA". In: *Science (New York, N.Y.)* 337.6099 (Sept. 7, 2012), pp. 1190–1195. ISSN: 0036-8075. DOI: 10.1126/science.1222794. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771521/ (visited on 04/23/2021).

[2] GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, et al. "Genetic effects on gene expression across human tissues". In: *Nature* 550.7675 (2017), pp. 204–213. ISSN: 1476-4687. DOI: 10.1038/nature24277.

[3] The GTEx Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (Sept. 11, 2020). Publisher: American Association for the Advancement of Science Section: Research Article, pp. 1318–1330. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaz1776. URL: https://science.sciencemag.org/content/369/6509/1318 (visited on 11/04/2020).

[4] Yuan He, Surya B. Chhetri, et al. "sn-spMF: matrix factorization informs tissue-specific genetic regulation of gene expression". In: *Genome Biology* 21.1 (Sept. 11, 2020), p. 235. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02129-6. URL: https://doi.org/10.1186/s13059-020-02129-6 (visited on 11/04/2020).

[5] Sarah M. Urbut, Gao Wang, et al. "Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions". en. In: *Nature Genetics* 51.1 (Jan. 2019), pp. 187–195. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0268-8. URL: https://www.nature.com/articles/s41588-018-0268-8 (visited on 09/11/2020).

[6] Benjamin P. Fairfax, Peter Humburg, et al. "Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression". In: *Science* 343.6175 (Mar. 7, 2014). Publisher: American Association for the Advancement of Science Section: Research Article. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1246949. URL: https://science.sciencemag.org/content/343/6175/1246949 (visited on 11/04/2020).

[7] Brunilda Balliu, Matthew Durrant, et al. "Genetic regulation of gene expression and splicing during a 10-year period of human aging". In: *Genome Biology* 20.1 (Nov. 2019), p. 230. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1840-y. URL: https://doi.org/10.1186/s13059-019-1840-y (visited on 09/10/2020).

[8] Eric R. Gamazon, Ayellet V. Segrè, et al. "Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation". In: *Nature Genetics* 50.7 (July 2018), pp. 956–967. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0154-4. URL: https://www.nature.com/articles/s41588-018-0154-4 (visited on 04/26/2021).

[9] Sarah Kim-Hellmuth, François Aguet, et al. "Cell type–specific genetic regulation of gene expression across human tissues". In: *Science* 369.6509 (Sept. 11, 2020). ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaz8528. URL: https://science.sciencemag.org/content/369/6509/eaaz8528 (visited on 04/23/2021).

[10] Tiffany Amariuta, Kazuyoshi Ishigaki, et al. "Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements". In: *Nature Genetics* 52.12 (Dec. 2020), pp. 1346–1354. ISSN: 1546-1718. DOI: 10.1038/s41588-020-00740-8. URL: https://www.nature.com/articles/s41588-020-00740-8 (visited on 04/23/2021).

[11] Bryce van de Geijn, Hilary Finucane, et al. "Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability". In: *Human Molecular Genetics* 29.7 (May 8, 2020), pp. 1057–1067. ISSN: 0964-6906. DOI: 10.1093/hmg/ddz226. URL: https://doi.org/10.1093/hmg/ddz226 (visited on 04/23/2021).

[12] Benjamin D. Umans, Alexis Battle, and Yoav Gilad. "Where Are the Disease-Associated eQTLs?" In: *Trends in Genetics* 37.2 (Feb. 1, 2021), pp. 109–124. ISSN: 0168-9525. DOI: 10.1016/j.tig.2020.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0168952520302092 (visited on 04/26/2021).

[13] Mark N. Lee, Chun Ye, et al. "Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells". In: *Science (New York, N.Y.)* 343.6175 (Mar. 7, 2014),

p. 1246980. ISSN: 0036-8075. DOI: 10.1126/science.1246980. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4124741/ (visited on 11/04/2020).

[14] Towfique Raj, Katie Rothamel, et al. "Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes". In: *Science (New York, N.Y.)* 344.6183 (May 2, 2014), pp. 519–523. ISSN: 0036-8075. DOI: 10.1126/science.1249547. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4910825/ (visited on 11/04/2020).

[15] Alessandra Ferraro, Anna Morena D'Alise, et al. "Interindividual variation in human T regulatory cells". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.12 (Mar. 25, 2014), E1111–E1120. ISSN: 0027-8424. DOI: 10.1073/pnas.1401343111. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3970507/ (visited on 11/04/2020).

[16] Benjamin P Fairfax, Seiko Makino, et al. "GENETICS OF GENE EXPRESSION IN PRIMARY IMMUNE CELLS IDENTIFIES CELL-SPECIFIC MASTER REGULATORS AND ROLES OF HLA ALLELES". In: *Nature genetics* 44.5 (Mar. 25, 2012), pp. 502–510. ISSN: 1061-4036. DOI: 10.1038/ng.2205. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3437404/ (visited on 11/04/2020).

[17] Chun Jimmie Ye, Ting Feng, et al. "Intersection of population variation and autoimmunity genetics in human T cell activation". In: *Science (New York, N.Y.)* 345.6202 (Sept. 12, 2014), p. 1254665. ISSN: 0036-8075. DOI: 10.1126/science.1254665. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751028/ (visited on 11/04/2020).

[18] Luis B. Barreiro, Ludovic Tailleux, et al. "Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.4 (Jan. 24, 2012), pp. 1204–1209. ISSN: 0027-8424. DOI: 10.1073/pnas.1115761109. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3268270/ (visited on 11/04/2020).

[19] Chun Jimmie Ye, Jenny Chen, et al. "Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection". en. In: *Genome Research* 28.12 (Dec. 2018), pp. 1812–1825. ISSN: 1088-9051, 1549-5469. DOI:

10.1101/gr.240390.118. URL: https://genome.cshlp.org/content/28/12/1812 (visited on 04/23/2021).

[20] B. J. Strober, R. Elorbany, et al. "Dynamic genetic regulation of gene expression during cellular differentiation". In: *Science* 364.6447 (June 28, 2019), pp. 1287–1290. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaw0040. URL: https://science.sciencemag.org/content/364/6447/1287 (visited on 04/26/2021).

[21] Jérôme Carayol, Christian Chabert, et al. "Protein quantitative trait locus study in obesity during weight-loss identifies a leptin regulator". en. In: *Nature Communications* 8.1 (Dec. 2017), p. 2084. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02182-z. URL: https://www.nature.com/articles/s41467-017-02182-z (visited on 03/29/2021).

[22] Apolline Gallois, Joel Mefford, et al. "A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context". In: *Nature Communications* 10.1 (Oct. 2019), p. 4788. DOI: 10.1038/s41467-019-12703-7.

[23] Rachel Moore, Francesco Paolo Casale, et al. "A linear mixed-model approach to study multivariate gene–environment interactions". In: *Nature Genetics* 51.1 (2019), pp. 180–186.

[24] Peter Kraft, Yu-Chun Yen, et al. "Exploiting Gene-Environment Interaction to Detect Genetic Associations". In: *Human Heredity* 63.2 (2007). Publisher: Karger Publishers, pp. 111–119. ISSN: 0001-5652, 1423-0062. DOI: 10.1159/000099183. URL: https://www.karger.com/Article/FullText/99183 (visited on 11/04/2020).

[25] Eun Yong Kang, Buhm Han, et al. "Meta-Analysis Identifies Gene-by-Environment Interactions as Demonstrated in a Study of 4,965 Mice". In: *PLOS Genetics* 10.1 (Jan. 9, 2014). Publisher: Public Library of Science, e1004022. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004022. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004022 (visited on 11/04/2020).

[26] Fan Zhang, Kevin Wei, et al. "Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry". In: *Nature Immunology* 20.7 (July 2019), pp. 928–942. ISSN: 1529-2916. DOI: 10.1038/s41590-019-

0378-1. URL: https://www.nature.com/articles/s41590-019-0378-1 (visited on 06/16/2021).

[27] Marna McKenzie, Anjali K. Henders, et al. "Overlap of expression Quantitative Trait Loci (eQTL) in human brain and blood". In: *BMC Medical Genomics* 7.1 (June 3, 2014), p. 31. ISSN: 1755-8794. DOI: 10.1186/1755-8794-7-31. URL: https://doi.org/10.1186/1755-8794-7-31 (visited on 04/02/2021).

[28] Maria Gutierrez-Arcelus, Halit Ongen, et al. "Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing". In: *PLOS Genetics* 11.1 (Jan. 29, 2015). Publisher: Public Library of Science, e1004958. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004958. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004958 (visited on 04/02/2021).

[29] Kimberly R. Kukurba, Princy Parsana, et al. "Impact of the X Chromosome and sex on regulatory variation". In: *Genome Research* 26.6 (June 2016), pp. 768–777. ISSN: 1549-5469. DOI: 10.1101/gr.197897.115.

[30] Meritxell Oliva, Manuel Muñoz-Aguirre, et al. "The impact of sex on gene expression across human tissues". In: *Science* 369.6509 (Sept. 11, 2020). ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aba3066. URL: https://science.sciencemag.org/content/369/6509/eaba3066 (visited on 04/16/2021).

[31] Andrey A. Shabalin. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". eng. In: *Bioinformatics (Oxford, England)* 28.10 (May 2012), pp. 1353–1358. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts163.

[32] Buhm Han and Eleazar Eskin. "Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies". In: *The American Journal of Human Genetics* 88.5 (May 13, 2011). Publisher: Elsevier, pp. 586–598. ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2011.04.014. URL: https://www.cell.com/ajhg/abstract/S0002-9297(11)00155-8 (visited on 11/04/2020).

[33] Jae Hoon Sul, Buhm Han, et al. "Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches". In: *PLOS Genetics* 9.6 (June 13, 2013). Publisher: Public Library of Science, e1003491. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003491. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003491 (visited on 04/02/2021).

[34] Niek de Klein, Ellen A. Tsai, et al. "Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases". en. In: *bioRxiv* (Mar. 2021), p. 2021.03.01.433439. DOI: 10.1101/2021.03.01.433439.

[35] Tom R. Gaunt, Hashem A. Shihab, et al. "Systematic identification of genetic influences on methylation across the human life course". In: *Genome Biology* 17.1 (Mar. 2016), p. 61. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0926-z. URL: https://doi.org/10.1186/s13059-016-0926-z (visited on 09/10/2020).

[36] Yue Hu, Xi Xi, et al. "SCeQTL: an R package for identifying eQTL from single-cell parallel sequencing data". In: *BMC Bioinformatics* 21.1 (May 11, 2020), p. 184. ISSN: 1471-2105. DOI: 10.1186/s12859-020-3534-6. URL: https://doi.org/10.1186/s12859-020-3534-6 (visited on 11/04/2020).

[37] Monique G. P. van der Wijst, Harm Brugge, et al. "Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs". In: *Nature Genetics* 50.4 (Apr. 2018). Number: 4 Publisher: Nature Publishing Group, pp. 493–497. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0089-9. URL: https://www.nature.com/articles/s41588-018-0089-9 (visited on 11/04/2020).

[38] MGP van der Wijst, DH de Vries, et al. "The single-cell eQTLGen consortium". In: *eLife* 9 (Mar. 9, 2020). Ed. by Helena Pérez Valle, Peter Rodgers, et al. Publisher: eLife Sciences Publications, Ltd, e52155. ISSN: 2050-084X. DOI: 10.7554/eLife.52155. URL: https://doi.org/10.7554/eLife.52155 (visited on 11/04/2020).

[39] Lee J. Cronbach and Noreen Webb. "Between-class and within-class effects in a reported aptitude * treatment interaction: Reanalysis of a study by G. L. Anderson". In: *Journal of Educational Psychology* 67.6 (1975), pp. 717–724. DOI: 10.1037/0022-0663.67.6.717.

[40] R. J. Simes. "An improved Bonferroni procedure for multiple tests of significance". In: *Biometrika* 73.3 (1986), pp. 751–754. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/73.3.751. URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/73.3.751 (visited on 04/29/2021).

[41] Schwarzer G. "meta: An R package for meta-analysis". In: *R News* 7.7 (), pp. 40–45.

[42] Camila M. Lopes-Ramos, Cho-Yi Chen, et al. "Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues". In: *Cell Reports* 31.12 (June 23, 2020), p. 107795. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2020.107795. URL: https://www.sciencedirect.com/science/article/pii/S2211124720307762 (visited on 04/16/2021).

[43] Sven Heinz, Casey E. Romanoski, et al. "The selection and function of cell type-specific enhancers". In: *Nature reviews. Molecular cell biology* 16.3 (Mar. 2015), pp. 144–154. ISSN: 1471-0072. DOI: 10.1038/nrm3949. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517609/ (visited on 10/09/2020).

[44] D. S. Gross and W. T. Garrard. "Nuclease hypersensitive sites in chromatin". In: *Annual Review of Biochemistry* 57 (1988), pp. 159–197. ISSN: 0066-4154. DOI: 10.1146/annurev.bi.57.070188.001111.

[45] Igor Mandric, Jeremy Rotman, et al. "Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing". In: *Nature Communications* 11.1 (Dec. 2020), p. 3126. ISSN: 2041-1723. DOI: 10.1038/s41467-020-16857-7. URL: http://www.nature.com/articles/s41467-020-16857-7 (visited on 05/10/2021).

[46] Serghei Mangul, Harry Taegyun Yang, et al. "ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues". In: *Genome Biology* 19.1 (Feb. 15, 2018), p. 36. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1403-7. URL: https://doi.org/10.1186/s13059-018-1403-7 (visited on 05/10/2021).

[47] Cory C. Funk, Alex M. Casella, et al. "Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data across 27 Tissue Types". In: *Cell Reports* 32.7 (Aug. 18, 2020), p. 108029. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2020.108029.

[48] Ruma Banerjee and Cheng-Gang Zou. "Redox regulation and reaction mechanism of human cystathionine-beta-synthase: a PLP-dependent hemesensor protein". In: *Archives of Biochemistry and Biophysics* 433.1 (Jan. 1, 2005), pp. 144–156. ISSN: 0003-9861. DOI: 10.1016/j.abb.2004.08.037.

[49] J. P. Kraus, J. Oliveriusová, et al. "The human cystathionine beta-synthase (CBS) gene: complete sequence, alternative splicing, and polymorphisms". In: *Genomics* 52.3 (Sept. 15, 1998), pp. 312–324. ISSN: 0888-7543. DOI: 10.1006/geno.1998.5437.

[50] Edith Wilson Miles and Jan P. Kraus. "Cystathionine beta-synthase: structure, function, regulation, and location of homocystinuria-causing mutations". In: *The Journal of Biological Chemistry* 279.29 (July 16, 2004), pp. 29871–29874. ISSN: 0021-9258. DOI: 10.1074/jbc.R400005200.

[51] Angata T, Hayakawa T, et al. "Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates." In: *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* 20.12 (Oct. 1, 2006), pp. 1964–1973. ISSN: 0892-6638, 1530-6860. DOI: 10.1096/fj.06-5800com. URL: https://europepmc.org/article/MED/17012248 (visited on 04/02/2021).

[52] Erwin Goldberg, Edward M. Eddy, et al. "LDHC THE ULTIMATE TESTIS SPECIFIC GENE". In: *Journal of andrology* 31.1 (2010), pp. 86–94. ISSN: 0196-3635. DOI: 10.2164/jandrol.109.008367. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2915756/ (visited on 04/02/2021).

[53] Annalisa Buniello, Jacqueline A. L. MacArthur, et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47 (D1 Jan. 8, 2019), pp. D1005–D1012. ISSN: 1362-4962. DOI: 10.1093/nar/gky1120.

[54] Alex KleinJan. "The crucial role of dendritic cells in rhinitis". In: *Current Opinion in Allergy and Clinical Immunology* 11.1 (Feb. 2011), pp. 12–17. ISSN: 1528-4050. DOI: 10.1097/ACI.0b013e328342335f. URL: https://journals.lww.com/co-allergy/Fulltext/

2011/02000/The_crucial_role_of_dendritic_cells_in_rhinitis.4.aspx (visited on 06/16/2021).

[55] Wei Wang and Matthew Stephens. "Empirical Bayes Matrix Factorization". In: *arXiv:1802.06931 [stat]* (May 2, 2021). arXiv: 1802.06931. URL: http://arxiv.org/abs/1802.06931 (visited on 06/14/2021).

[56] Ali Pazokitoroudi, Yue Wu, et al. "Efficient variance components analysis across millions of genomes". In: *Nature Communications* 11.1 (Aug. 11, 2020), p. 4020. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17576-9. URL: https://www.nature.com/articles/s41467-020-17576-9 (visited on 06/15/2021).

[57] Claudia Giambartolomei, Jimmy Zhenli Liu, et al. "A Bayesian framework for multiple trait colocalization from summary association statistics". In: *Bioinformatics (Oxford, England)* 34.15 (Aug. 1, 2018), pp. 2538–2545. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty147.

[58] C. B. Peterson, M. Bogomolov, et al. "TreeQTL: hierarchical error control for eQTL findings". In: *Bioinformatics (Oxford, England)* 32.16 (Aug. 15, 2016), pp. 2556–2558. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw198.

[59] Gabriel E. Hoffman and Eric E. Schadt. "variancePartition: Interpreting drivers of variation in complex gene expression studies". In: *BMC Bioinformatics* 17 (483 2016). DOI: 10.1186/s12859-016-1323-z.

[60] Daniel Ho, Kosuke Imai, et al. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference". In: *Political Analysis* 15 (2007), pp. 199–236.

[61] Jian Yang, S. Hong Lee, et al. "GCTA: A Tool for Genome-wide Complex Trait Analysis". In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.